KU LEUVEN

ARENBERG DOCTORAL SCHOOL
Faculty of Science

# Computational study of HIV gp41 FP and Rev as novel antiretroviral targets

**Tom Venken**

Supervisor:
Prof. dr. M. De Maeyer

Co-supervisor:
Dr. A. Voet

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor in Science: Biochemistry and Biotechnology

December 2013

# Computational study of HIV gp41 FP and Rev as novel antiretroviral targets

**Tom VENKEN**

Supervisory Committee:
Prof. dr. H. Mizuno, chair
Prof. dr. M. De Maeyer, supervisor
Dr. A. Voet, co-supervisor
Prof. dr. D. Daelemans
Prof. dr. W. De Borggraeve
Prof. dr. J. Robben
dr. C. Boutton (Ablynx N.V.)

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor in Science: Biochemistry and Biotechnology

December 2013

D/2013/10.705/103
ISBN 978-90-8649-689-1

# Dankwoord

"You act just like a
scientist and i like a woman"

Warehouse - Dead Man Ray

Dit is waarschijnlijk het onderdeel waar ik het meeste aandacht aan zou moeten schenken - want laten we eerlijk zijn - dit deel zal in de praktijk ook het meest worden gelezen. Er zijn dan ook meerdere personen die ik graag zou willen bedanken, waarbij ik mij bij deze al verontschuldig voor al degenen die ik toevallig over het hoofd zou zien.

Allereerst zou ik mijn promotor prof. dr. Marc De Maeyer willen bedanken. Marc, u stond altijd klaar om mij met raad en daad te helpen, of het nu over de algemene lijnen van het project ging of de piepkleine details. Uw deur stond altijd voor mij open, of er papers verbeterd moesten worden, maar ook gewoon om het over de kleine dinges des levens te hebben ... zoals waar je bijvoorbeeld appelen tot appelsap kan laten persen. Uw steun was onvoorwaardelijk, daarom bij deze een onvoorwaardelijk dankuwel!

Natuurlijk wil ik ook dank betuigen aan mijn co-promotor, dagelijkse mentor en sensei dr. Arnout Voet. Het is een lange wetenschappelijke reis van 4 jaar geweest en je was een goede gids! En dit zowel wetenschappelijk als als reisbegeleider, zowel in soms zeer hoge kerktorens (voor mij) als mogelijk iets te lage grottempels (voor jou). En ja, ook al heb ik soms gevloekt en op mijn tanden moeten bijten, bedankt om zo achter mijn veren te zitten, anders had ik hier nu niet gestaan ;)

I also like to thank the jury members for all your suggestions: prof. dr. Mizuno, prof. dr. Daelemans prof. dr. De Borggraeve, prof. dr. Robben and dr. Boutton.

Dank gaat uiteraard ook uit naar al de leden van het Biomol labo. Abel, nogmaals bedankt om me tijdens mijn masterthesis mijn eerste stappen in de wetenschap te laten zetten. De cruncher is toch een stuk meer gecrasht sinds jij weg was. Mieke en Renée, bedankt voor de vele babbels en de gezellige sfeer. Het werd een stuk stiller toen jullie ons labo verlieten voor andere oorden, maar gelukkig zijn er ook meerdere personen ons labo komen vergezellen. Joren, je was een zeer fijne collega en we hebben heel wat wateren doorzeild, zowel Tip3p als het IJsselmeer. Op een dag sta je hier ook en zal ik je trots aanspreken met dr. Scientific Vector Lad ;) Xiaoyu, good luck with your PhD and I hope you can become dr. QSAR girl one day! Wim G. en Dries, bedankt dat ik meerdere keren jullie begeleider mocht zijn en nog veel succes. Bij deze wil ik mij ook nog verontschuldigen dat ik jullie met

een gigantisch spinnenweb van paperclips heb opgezadeld. Ook dank aan mijn andere studenten Marlies, Michael en Sven. Tim, we hebben maar even voor een TNT spraakverwarring gezocht, maar bedankt voor de fijne tijd. Ik hoop dat je nog goedgemutst naar Blankenberge kan reizen zonder een bepaalde muziektune in je hoofd te krijgen. Jonas, het was altijd een fijn weerzien op een labotrip of een BBQ! En Wim V.D.B., bedankt om extra gezelligheid in onze bureau te brengen en mijn excuses voor het veelvoudig verstellen van je bureaustoel.

Natuurlijk mag ik alle andere (al dan niet voormalige) collega's van onze verdieping niet vergeten: Annelies, Amin, Bendy, Benjamien, Charlotte, Doortje, Elke, Gosia, Guoli, Herlinde, Jacek, Jelle, Jeroen, Jessika, Louis, Manli, Peter, Sangeetha, Sam, Sarah S., Sarah T., Vanni, Vincent, Yves E. en Werner. Ook dank aan de jongens en meisjes van de tweede verdieping: Amber, Annemie, Ching-Wen, Giovanni, Iris, Kasia, Kherim, Lorenzo, Mariem, Misha, Veerle, Wouter en Yves P. Els en Karin, bedankt om al mijn administratieve zorgen met de glimlach op te lossen!

There are evidently a number of people that I would like to thank for their essential contribution to this work. Jan, Frank, Daniela and Petra, thank you for your collaboration on the VIRIP project. Special thanks to Kashif as we started a collaboration just by meeting on a conference and it was very nice working together, mostly by Skype than by actually seeing each other. I also like to thank Gianni for his useful advice and critical eye, and of course the volunteers of gpu-grid. Thanks to Guy, Sylvie and Mohammadreza for guiding me in the intriguing world of beer biochemistry. And last but not least, zou ik naast Dirk natuurlijk ook Eline, Els, Maarten en Thomas van het Rega instituut willen bedanken voor de fijne samenwerking en leerrijke discussies over Rev.

Bijzondere dank aan het IWT - SBO - PharmAbs en de GOA Multicentre quantum chemistry voor de financiële steun. Ook dank aan het FWO, die me hebben toegelaten om enkele maanden in een buitenlands labo te werken. Special thanks to Kam for welcoming me at your laboratory. Thanks of course as well to Ashutosh, David, Francois, Kamlesh, Muhammad, Rojan, Taeho and Xiao Yin. Perhaps we can play table billiard or go to the wataminchi again some day!

Ook een welgemeend dankuwel aan mijn ouders voor al de steun de afgelopen jaren en aan mijn zus die meerdere keren teksten van mij heeft nagelezen als experte in Engels-Nederlands. Anthony, bedankt dat je zo goed voor mijn zus zorgt. En dank ook aan mijn schoonouders, mijn excuses dat ik jullie dochter soms iets teveel op sleeptouw heb genomen.

En als laatste maar als meeste bedankt... Liefste Niky, bij deze wil ik je heel erg bedanken voor de afgelopen vier jaar en me verontschuldigen als ik niet altijd genoeg tijd voor je had, maar weer net iets teveel tijd op het verkeerde moment aan dat doctoraat zat te besteden. Je hebt me altijd gesteund en je hebt niet geaarzeld om me zelfs tot in andere kanten van de wereld te volgen. Met de bus naar het land van de goulash, met het vliegtuig naar het land van de rijzende zon, ik kon altijd op je rekenen. Zonder jou was het niet gelukt. Ik zie je graag!

# Abstract

The human immunodeficiency virus (HIV) is the causative agent of acquired immunodeficiency syndrome (AIDS), a disease that to this day has resulted into more than 25 million deaths in the world. Despite the development of multiple antiretroviral drugs in the last twenty years, no effective vaccine or cure is currently available. Developing new drug classes with alternative mode of actions is a promising and innovative approach. Before the start of this PhD, a peptide inhibitor called VIRIP was discovered that binds to the N-terminal fusion peptide (FP) of gp41, an extracellular viral protein essential for fusion with human host cells and hence viral infectivity. A combined molecular dynamics (MD) simulation and specifically optimised binding free energy calculation approach was used to analyse the interactions between gp41 FP and the multiple VIRIP derivatives. Enhanced VIRIP derivatives were suggested and a selection was subsequently tested *in cellulo*. While the FP as target has been studied extensively in membrane and solution environments, its structure remained largely ambiguous to this day. Therefore, another goal was the characterisation of the FP in solution environments using atomistic MD simulations. Finally, another interesting antiviral target emerged during my research, namely the multimerisation process of the Rev protein. Interactions between individual Rev monomers were studied using the MD and binding free energy calculation protocol. The hot spot residues in each binding interface were revealed and their energy values were found to be in correlation with previous experimental measurements. It is expected that this information may guide the development of novel Rev specific antiretrovirals.

# Beknopte samenvatting

Het humaan immuundeficiëntievirus (HIV) is verantwoordelijk voor het veroorzaken van het verworven immuundeficiëntiesyndroom (AIDS). Deze ziekte heeft tot meer dan 25 miljoen doden geleid tot op de dag van vandaag. Er is nog steeds geen effectief vaccin of geneesmiddel ontdekt voor deze ziekte, ondanks de ontwikkeling van meerdere antiretrovirale middelen in de laatste twintig jaar. De ontwikkeling van nieuwe middelen met alternatieve mechanismes is een interessante en vernieuwende aanpak. Voor het begin van dit onderzoek is een peptide genaamd VIRIP ontdekt. Dat peptide bindt aan het N-terminaal fusiepeptide (FP) van gp41, dat essentieel is voor fusie van virale partikels met gastheercellen en bijgevolg van virale infectiviteit. Een combinatie van moleculaire dynamica (MD) simulaties en specifiek geoptimaliseerde bindingsenergie-methoden werd gebruikt voor de analyse van de interacties tussen gp41 FP en VIRIP-derivaten. Verbeterde VIRIP-derivaten werden voorgesteld en een selectie werd in cellen getest. Hoewel het FP als doelwit uitgebreid is bestudeerd in membraan en oplossing, is de structuur tot op de dag van vandaag dubbelzinnig. Daarom was een ander doel van dit onderzoek de karakterisatie van het FP in oplossing met behulp van MD-simulaties. Ten slotte hebben we een ander interessant antiviraal doelwit onderzocht, namelijk de multimerisatie van het Rev eiwit. Interacties tussen individuele Rev monomeren werden bestudeerd door middel van het zonet vermelde MD- en bindingsenergie-protocol. De belangrijkste residu's in elke bindingsplaats werden opgehelderd en de energiewaarden correleerden met vroegere experimentele metingen. Het is te verwachten dat deze informatie de ontwikkeling van nieuwe Rev-specifieke antiretrovirale middelen kan stimuleren.

# Abbreviations

| | |
|---|---|
| $\alpha$ interface | The dimerisation Rev interface |
| $\beta$ interface | The higher order multimerisation Rev interface |
| aa | Amino Acid |
| ACE | Acetyl Group |
| AFM | Atomic Force Microscopy |
| AIDS | Acquired ImmunoDeficiency Virus |
| AMBER | Assisted Model Building with Energy Refinement |
| ARM | Arginine Rich Motif |
| AUC | Area Under the Curve |
| AZT | azidothymidine / zidovudine |
| BAR | Bennet's Acceptance Ratio |
| BIV | Bovine Immunodeficiency Virus |
| bp | Base Pair |
| CADD | Computer-Aided Drug Design |
| CCR5 | C-C chemokine Receptor type 5 |
| CD | Circular Dichroism |
| CHARMM | Chemistry at HARvard Molecular Mechanics |
| CPU | Central Processing Unit |
| CXCR4 | C-X-C chemokine Receptor type 4 |
| DNA | DeoxyRibonucleic Acid |
| DSSP | Define Secondary Structure of Proteins |
| EI | Entry Inhibitor |
| ELISA | Enzyme-Linked ImmunoSorbent Assay |
| EPR | Electron Paramagnetic Resonance Spectroscopy |
| FEP | Free Energy of Perturbation |
| FI | Fusion Inhibitor |
| FIV | Feline Immunodeficiency Virus |
| FP | Fusion Peptide |
| FP23 | The FP sequence with the highest prevalence |
| FPPR | Fusion Peptide Proximal Region |
| FRET | Förster Resonance Energy Transfer |
| FTIR | Fourier Transform Infrared Spectroscopy |

| | |
|---|---|
| GB | Generalised Born |
| gp | GlycoProtein |
| GPU | Graphical Processing Unit |
| GROMACS | GROningen Machine for Chemical Simulations |
| GROMOS | GROningen MOlecular Simulation |
| HAART | Highly Active AntiRetroviral Therapy |
| HGV | Hepatitis G Virus |
| HIV | Human Immunodeficiency Virus |
| HPC | High Performance Computing |
| HR | Heptad Repeat |
| IDP | Intrinsically Disordered Proteins |
| INI | Integrase Inhibitor |
| ITC | Isothermal Titration Calorimetry |
| LIE | Linear Interaction Energy |
| LINCS | Linear Constraint Solver |
| LJ | Lennard Jones |
| LTR | Long Terminal Repeats |
| M1 | The first multimerisation domain of Rev |
| M2 | The second multimerisation domain of Rev |
| MD | Molecular Dynamics |
| MM | Molecular Mechanics |
| MM-GBSA | Molecular Mechanics/Generalised-Born Surface Area |
| MM-PBSA | Molecular Mechanics/Poisson-Boltzmann Surface Area |
| MOE | Molecular Operating Environment |
| MPER | Membrane Proximal External Region |
| MSM | Markov State Model |
| NAMD | Not just Another Molecular Dynamics program |
| $Nb_{190}$ | A llama nanobody inhibiting Rev multimerisation |
| NES | Nuclear Export Signal |
| NIS | Nuclear Entry Inhibitory Signal |
| NLS | Nuclear Localisation Signal |
| NM | Normal Mode |
| NME | N-methyl Amide Group |
| NMR | Nuclear Magnetic Resonance |
| NNRTI | Non-Nucleoside Reverse-Transcriptase Inhibitor |
| NPT | Ensemble with constant number of particles, pressure and temperature |
| NRTI | Nucleoside Reverse-Transcriptase Inhibitor |

| | |
|---|---|
| NVE | Ensemble with constant number of particles, volume and energy |
| NVT | Ensemble with constant number of particles, volume and temperature |
| OPLS | Optimised Potentials for Liquid Simulations |
| PB | Poisson Boltzmann |
| PBC | Periodic Boundary Conditions |
| PDB | RCSB Protein Databank |
| PI | Protease Inhibitor |
| PIC | Pre-Integration Complex |
| PKI | Protein Kinase Inhibitor |
| PME | Particle Mesh Ewald |
| PMF | Potential of Mean Force |
| PPI | Protein Protein Interaction |
| QM | Quantum Mechanics |
| Rev | Regulator of Expression of Virion proteins |
| RMSD | Root Mean Square Deviation |
| RMSF | Root Mean Square Fluctuations |
| RNA | RiboNucleic Acid |
| ROC | Receiver Operator Characteristics |
| RRE | Rev Responsive Element |
| RTI | Reverse-Transcriptase Inhibitor |
| SA | Surface Area |
| SAR | Structure Activity Relationship |
| SASA | Solvent Accessible Surface Area |
| SIV | Simian Immunodeficiency Virus |
| SMD | Steered Molecular Dynamics |
| SMPPII | Small Molecule Protein-Protein Interaction Inhibitor |
| SPR | Surface Plasmon Resonance |
| T20 | Enfuvirtide / Fuzeon |
| TI | Thermodynamic Integration |
| VIR-165 | An optimised VIRIP derivative containing a disulphide bond |
| VIR-576 | A dimeric optimised VIRIP derivative |
| VIRIP | VIRus INhibitory Peptide |
| wt | wild type |

# List of Symbols

| | | |
|---|---|---|
| Å | Ångström | $10^{-10}$ m |
| $\Delta\Delta G$ | Relative binding free energy | kcal mol$^{-1}$ |
| $\Delta G$ | Absolute binding free energy | kcal mol$^{-1}$ |
| $\epsilon$ | Dielectric constant | |
| $\epsilon_p$ | Protein dielectric constant | |
| $\epsilon_w$ | Water dielectric constant | |
| $\kappa^2$ | Debye-Hückel constant | kg m$^{-3}$ |
| $\lambda$ | Eigenvector | |
| $\mu$s | microsecond | $10^{-6}$ s |
| $\nabla$ | Differential operator | |
| $\omega$ | Covariance matrix | |
| $\psi$ | Electrostatic potential | V |
| $\rho$ | Density | kg m$^{-3}$ |
| $\sigma$ | Frequency | Hz |
| $a$ | Acceleration of a particle | m s$^{-2}$ |
| $E_b$ | Bonded force field terms | kcal mol$^{-1}$ |
| $E_{angle}$ | The angle bending term in a force field | kcal mol$^{-1}$ |
| $E_{bond}$ | The bond stretching term in a force field | kcal mol$^{-1}$ |
| $E_{coul}$ | The electrostatic contribution in a force field | kcal mol$^{-1}$ |
| $E_{dihedral}$ | The dihedral torque term in a force field | kcal mol$^{-1}$ |
| $E_{nb}$ | Non-bonded force field terms | kcal mol$^{-1}$ |
| $E_{vdw}$ | The van der Waals term in a force field | kcal mol$^{-1}$ |
| $F$ | Force | N |
| $G$ | Gibbs Free energy | kcal mol$^{-1}$ |
| $G^{\circ}_{bind}$ | Binding free energy under standard conditions | kcal mol$^{-1}$ |
| $G_{ele-tot}$ | Total electrostatic free energy | kcal mol$^{-1}$ |
| $G_{GB}$ | Polar solvation free energy derived by GB | kcal mol$^{-1}$ |
| $G_{hyd-tot}$ | Total hydrophobic free energy | kcal mol$^{-1}$ |
| $G_{MM}$ | Molecular Mechanics free energy | kcal mol$^{-1}$ |

| | | |
|---|---|---|
| $G_{PB}$ | Polar solvation free energy derived by PB | kcal $mol^{-1}$ |
| $G_{SA}$ | Apolar solvation free energy | kcal $mol^{-1}$ |
| $G_{solv}$ | Solvation free energy | kcal $mol^{-1}$ |
| $G_{sub-tot}$ | Subtotal free energy | kcal $mol^{-1}$ |
| $G_{tot}$ | Total free energy | kcal $mol^{-1}$ |
| $H$ | Enthalpy | kcal $mol^{-1}$ |
| $IC_{50}$ | The half maximal inhibitory concentration | M |
| $K_a$ | Association constant | $M^{-1}$ |
| $k_b$ | Boltzmann constant | 1.38 x $10^{-23}$ J $K^{-1}$ |
| $K_d$ | Dissociation constant | M |
| $k_{on}$ | Off-rate constant | $s^{-1}$ (for first-order kinetics) |
| $k_{on}$ | On-rate constant | $s^{-1}$ (for first-order kinetics) |
| $m$ | Mass of a particle | D |
| $N$ | Number of particles | |
| $p$ | Pressure | bar |
| $q$ | Charge | C |
| $R$ | Ideal gas constant | 1.987 cal $K^{-1}$ $mol^{-1}$ |
| $S$ | Entropy | kcal $mol^{-1}$ $K^{-1}$ |
| $S_{conf}$ | Conformational/configurational entropy | kcal $mol^{-1}$ $K^{-1}$ |
| $S_{solv}$ | Entropic solvation free energy | kcal $mol^{-1}$ $K^{-1}$ |
| $T$ | Temperature | T |
| $t$ | Time | s |
| $U$ | Total potential energy | kcal $mol^{-1}$ |
| $V$ | Volume | $m^3$ |
| $v$ | Velocity of a particle | m $s^{-1}$ |
| fs | femtosecond | $10^{-15}$ s |
| nm | nanometer | $10^{-9}$ m |
| ns | nanosecond | $10^{-9}$ s |
| ps | picosecond | $10^{-12}$ s |

# Contents

# Part I

# Aims

The human immunodeficiency virus (HIV) is the causative agent of acquired immunodeficiency syndrome (AIDS). This disease has resulted into more than 25 million deaths in the world to this day. Despite the development of multiple antiretroviral drugs in the last twenty years, no effective vaccine or cure is currently available. Antiretroviral treatment is thwarted by an increased emergence of resistance in HIV-infected patients due to the high mutation and replication rate of HIV. This results in elevated genetic variability making current antiretroviral drugs less effective. Resistant effects can be limited by improving current antiretroviral drug classes even further. Alternatively, developing new drug classes with alternative mode of actions would be a more promising and innovative approach. Preferably, those new drug classes should contain very high resistance barriers to avoid accumulation of mutations in viral proteins. The inclusion of these novel drugs targeting other steps in the viral replication cycle is becoming attractive for the effectiveness of antiretroviral therapies.

At the start of this PhD, a peptidic inhibitor called VIRIP (VIRus INhibitory Peptide) has been reported by the group of prof. dr. Frank Kirchhoff and prof. dr. Jan Münch at the Institute of Molecular Virology, Ulm University Medical Center, Germany. This peptide binds to the N-terminal fusion peptide (FP) of gp41, an extracellular viral protein essential for fusion with human host cells and hence viral infectivity. The gp41 FP is an interesting novel antiretroviral target due to its highly conserved nature and importance during viral replication. Multiple VIRIP derivatives have been developed, but the mode of action of VIRIP remained largely unclear. Therefore, the first objective of this PhD is the **analysis of the interactions between gp41 FP and the multiple VIRIP derivatives**. To this end, a combined molecular dynamics (MD) simulation and a specifically optimised binding free energy calculation approach was used. Based on this analysis, another aim was to **suggest VIRIP derivatives with improved antiretroviral activity** to test a selection *in cellulo* by the group of prof. dr. Frank Kirchhoff and prof. dr. Jan Münch.

While the FP as target has been studied extensively in both membrane and solution environments, its structure remains largely ambiguous to this day. Therefore, a second objective is to **characterise the FP in solution using atomistic MD simulations**. These calculations were performed in collaboration with dr. Kashif Sadiq and dr. Gianni De Fabritiis of the Computational Biophysics Laboratory, Universitat Pompeu Fabra, Barcelona. These studies do not only give an atomistic glimpse of the conformational distribution of the FP, they can also suggest conformations that would be interesting to target by fusion inhibitors.

During the course of my research, another interesting antiviral target emerged: the multimerisation process of the Rev protein. This protein mediates nuclear export of viral mRNA and as a consequence is important for the onset of the late HIV replication cycle. Multiple Rev monomers multimerise and form a functional multimeric complex with viral RNA. This multimerisation event is indispensable for the Rev function, as such development of small molecule inhibitors targeting the multimerisation process would be a novel antiretroviral approach. Crystal structures of dimeric and tetrameric Rev structures were recently resolved, but unfortunately did not allow for a full understanding of the multimerisation process. Therefore, a third objective is to **study the interactions between individual Rev monomers** using the previously mentioned combined MD and binding free energy calculation protocol.

Finally, a last objective is to utilise the information from the multimerisation interaction study to conduct a **virtual screening** study using a 3D pharmacophore model for the discovery of **novel Rev multimerisation inhibitors**. This study is performed in collaboration with the group of prof. dr. Dirk Daelemans of the Rega Institute, KU Leuven, who has developed and optimised various *in vitro* and *in vivo* methods to study the multimerisation of Rev. Note that the results of this study are not included in this thesis for protection of intellectual property.

The aims of the thesis can be summarized as follows and are investigated in the listed result chapters:

| Chapter | Aim |
|---|---|
| 4 | Analyse interactions between gp41 FP and VIRIP derivatives |
| 4 | Suggest VIRIP derivatives with improved antiretroviral activity |
| 5 | Characterise the FP in solution using atomistic MD simulations |
| 6 | Study the interactions between individual Rev monomers |
| / | Virtual screening of novel Rev multimerisation inhibitors |

# Part II

# Introduction

In this part, we will introduce a background of the human immunodeficiency virus and provide some general concepts applied in this thesis. The part is divided in the following chapters.

**Chapter 1 | HIV: the role of gp41 FP and Rev**

This chapter outlines the biological background of the doctoral research. HIV-1 and its structure and replication cycle are introduced in detail. We explain the current state of research of the two targets in this thesis, the gp41 fusion peptide and the regulatory Rev protein.

**Chapter 2 | Molecular dynamics simulations**

Here, we introduce MD simulations as a valuable tool to investigate molecules of interest. A perspective of the method compared to experimental methods is presented. We explain the theoretical concepts of MD simulations in detail and offer some limitations and challenges in the field.

**Chapter 3 | Binding free energy calculations**

The concepts in the previous chapter can be implemented to estimate binding affinities between biomolecules, an important task in drug development. An overview of the current binding free energy methods is presented, where we shortly discuss their strengths and weaknesses. The method used in this thesis, MM-PB/GBSA, is summarised in detail. Furthermore, we present a number of applications and underline a number of considerations of the method.

# Chapter 1

## HIV: the role of gp41 FP and Rev

"Take Your Protein Pills And
Put Your Helmet On"

Space Oddity, David Bowie

### 1.1 Human Immunodeficiency Virus: An introduction

In 1981, an unknown illness was reported in previously healthy men in Los Angeles. Patients were characterised by an unusual amount of opportunistic infections and rare malignant tumours, such as Kaposi's sarcoma and Burkitt's lymphoma. The disease was found to spread through sexual transmission or blood transfusion and was called acquired immunodeficiency syndrome (AIDS) before the causative agent was discovered. Many different theories were proposed as the cause of the peculiar disease, but it took until 1983 before the human immunodeficiency virus (HIV) was identified as the etiological agent of AIDS [1, 2]. Until today, AIDS has evolved into one of the greatest pandemics caused by a previously unrevealed infectious agent in modern history, with millions of people infected around the world. AIDS is not only a disease. It has strong economic implications, especially in certain Sub-Saharan African countries where mostly young people are affected; resulting in a significantly reduced life expectancy and increased poverty levels.

HIV is a retrovirus, belonging to the class of the lentiviridae. *Lente* is Latin for "slow" as exemplified by the long incubation period of the virus in the host. HIV is characterised by a complex immunology [3] and is closely related to other mammalian retroviruses such as simian immunodeficiency virus (SIV), feline immunodeficiency virus (FIV) and bovine immunodeficiency virus (BIV). Two different types can be distinguished: HIV-1 (a global variant) and HIV-2 (mainly restricted to Western-Africa). HIV-1 is the most virulent and widespread variant and it is therefore the most studied type. The viral particles target white blood cells

harbouring CD4-receptors, such as T helper cells (these are lymphocytes important in the adaptive immune system by release of cytokines) and macrophages (these are lymphocytes part of the innate immune system that digest pathogens and remove cellular waste). Upon infection, HIV subdues the host cell machinery and the viral genetic material is irreversibly integrated in human cells. HIV infection leads to a significant decrease in cellular immunity due to T cell degradation, thereby abrogating the immune system response. That decrease ultimately results in the disease we call AIDS.

In 1987, the first antiretroviral drug zidovudine (AZT) was approved for the treatment of AIDS and HIV infection [4]. Until now, more than 25 antiretroviral drugs have been licensed. Five different HIV drug classes can be distinguished, each one targeting different steps in the replication cycle (see Figure 1.1): viral entry inhibitors (EIs) including fusion inhibitors (FIs) and coreceptor antagonists, reverse transcriptase inhibitors (RTIs) such as nucleoside analogue (NRTIs) and non-nucleoside analogue inhibitors (NNRTIs), integrase inhibitors (INIs) and protease inhibitors (PIs). An overview of all the currently licensed drugs can be found in a review article by De Clercq [5].

Today, HIV-positive patients are treated with highly active antiretroviral therapy (HAART) regimens, which contain a combination of three or four drugs that target at least two different viral proteins. While HAART can successfully suppress the viral loads for multiple years in HIV-positive patients, the virus is never entirely eradicated and remains in the human body in latent reservoir cells [6]. The disease can thus be treated but it can never be entirely cured with the current HAART regimens. There are, however, a few notable exceptions. Recently, a patient who received haematopoietic stem cell transplantations to treat leukaemia was found to no longer harbour detectable HIV viraemia, even after quitting antiretroviral therapy [7, 8]. A similar result was found in two other HIV-positive patients who were given bone marrow transplants to treat cancer [9, 10]. Thus, except for the special case of replacing stem cells in the bone marrow, the virus is never completely eliminated from the body.

While those latest experiments suggest that HIV infection might be curable one day, stem cell transplantation is unfortunately not a realistic treatment option due to the high risks and expensive procedures involved. Hence, there is still a profound need to develop novel antiretroviral drugs effective against highly resistant virus strains to complement or extend current HAART regimens. Moreover, while the number of AIDS-related deaths has decreased remarkably from the success of antiretroviral therapies, the number of people living with HIV-1 continues to increase [11]. Current antiretroviral drugs also have a significant number of shortcomings, such as toxicity issues, high manufacturing or licensing costs

and emergence of side effects. As such, a thorough biological and structural understanding of HIV is required to develop novel antiretroviral drugs or possible vaccines. This effort depends on collaboration between researchers in academia, governmental organisations and the pharmaceutical industry. In this chapter, we aim to provide an overview of the structure and genome of HIV, the HIV infection cycle and a summary of the targets in this thesis, gp41 FP and Rev.

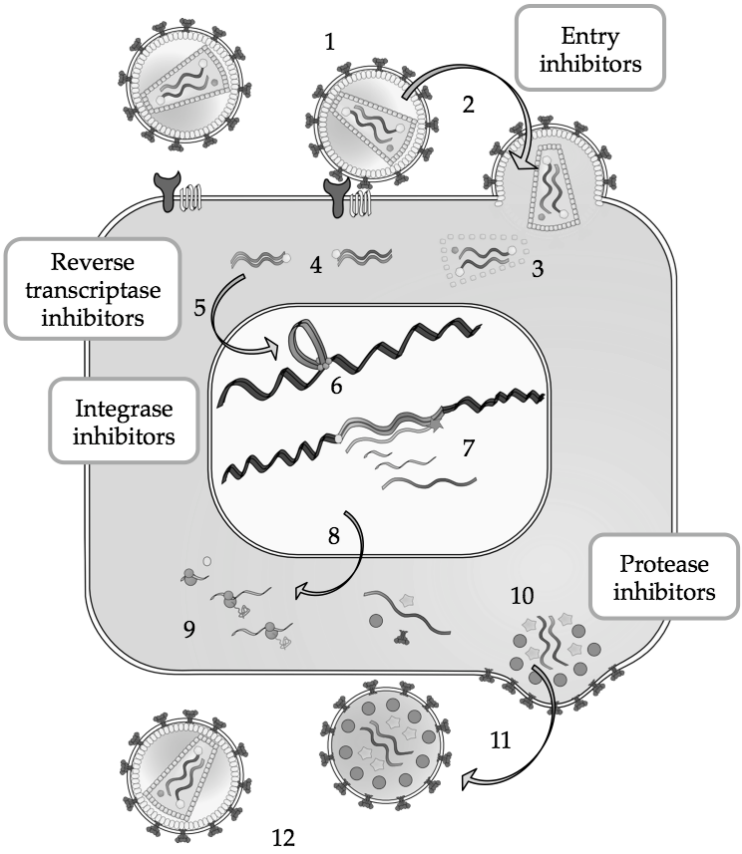


Figure 1.1: **Replication cycle of HIV.** After attachment of a HIV virion (1) and subsequent fusion of viral and cellular membranes (2), the viral genetic material enters the host cell. The HIV genome, consisting of two single-stranded RNA copies, is converted into double-stranded DNA by reverse transcriptase (4). Next, the viral DNA is imported into the nucleus (5) and integrated into the host genome by integrase (6). Multiple viral copies are transcribed (7) and are after export to the cytoplasm (8) translated into new viral proteins (9). A novel HIV particle is formed (10) and buds out of the host cell (11). This new HIV virion becomes infective after maturation of viral proteins by the protease enzyme (12). The boxes indicate the current developed antiretroviral inhibitors licensed on the clinical market. In this PhD thesis, we target the fusion step (2) and the export of newly transcribed RNA strands from the nucleus to the cytoplasm by the Rev protein (8). Figure adapted from Voet *et al.* [12].

## 1.2 Structure and genome of HIV

HIV is a retrovirus, that is, an enveloped virus with genetic material stored in a RNA genome. HIV contains two identical single-stranded positive-sense RNA copies of 9.7 kb. These RNA strands are converted (reverse-transcribed) into DNA by the enzyme reverse transcriptase during the replication cycle. The generated DNA copy can subsequently be integrated in the target cell genome by another viral enzyme called integrase and host co-factor proteins. The DNA chromosome integration depends on long terminal repeats (LTRs) at the ends of the RNA genome, which are also responsible for gene expression after integration.

The HIV genome contains a total of nine genes. Three genes are shared with other retroviruses and are essential for the viral lifecycle: *gag* (group specific antigen), *pol* (polymerase) and *env* (envelope). These genes are translated into matrix/capsid proteins (necessary for stabilisation of the virion), enzymatic proteins (necessary for replication) and envelope proteins (necessary for viral attachment and fusion to host cells). In contrast with other retroviruses, a lentivirus such as HIV contains additional regulatory genes. *rev* (regulator of viral proteins) and *tat* (transactivator) code for regulatory proteins, while *nef*, *vif*, *vpu* (viral protein U) and *vpr* (viral protein R) are translated into accessory proteins. The organisation of the viral genome can be found in Figure 1.2.

The HIV virion is approximately 100 to 120 nm in diameter, which is relatively large for a virus. The outer shell, the viral envelope, comprises a mixture of host cell membrane and viral membrane proteins embedded in a phospholipid bilayer. Inside the envelope, matrix proteins are found that stabilise the virion by attachment to the inner lipid membrane. Next, a cone-shaped nuclear shell made of capsid proteins is found. Finally, the two RNA strands and the replication enzymes such as reverse transcriptase, integrase and protease are contained inside the core.

## 1.3 HIV replication cycle

The infection cycle of HIV-1 commences when the viral envelope glycoproteins attach to human host cells (see Figure 1.1) [14, 15]. The *env* gene codes for the gp160 protein, which is cleaved by the human protease furin into two non-covalently linked trimeric proteins, gp120 and gp41. These two proteins form a trimer-of-heterodimers that attaches to the target cell membrane in multiple steps. First, gp120 interacts with a CD4 receptor on a host cell. This event triggers a conformational change in gp120, allowing binding with a chemokine coreceptor,

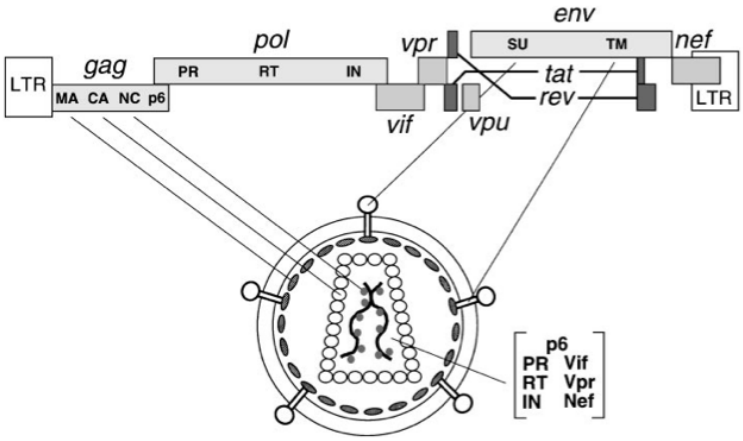


Figure 1.2: **Illustration of the HIV-genome and virion**. Figure from Frankel and Young [13]. See the text for details of the abbreviations.

CCR5 (for R5-tropic strains) or CXCR4 (for X4-tropic strains), thereby reducing the distance between gp120 and the host cell surface. Now that both surfaces lie in close proximity, gp41 undergoes a conformational change as well and inserts its N-terminal fragment, called the fusion peptide (FP), into the host cell membrane. The N- and C-helices of gp41 form an exceptionally stable six-helix bundle after the insertion event. This oligomerisation is energetically beneficial and multiple gp41 molecules cooperate to allow fusion of both viral and cell membrane. Finally, HIV genetic material and viral enzymes are released into the host cell.

After the uncoating of the HIV particle, reverse transcriptase uses the single-stranded RNA genome to produce single-stranded DNA, which is subsequently converted into a double-stranded DNA copy of the viral genome. Next, many viral and host proteins join forces to form a so-called pre-integration complex (PIC), which is transported into the host cell nucleus afterwards. The reverse transcriptase is tremendously error-prone, lacking a mechanism to correct mistakes while reproducing the genome. Hence, the enzyme has a high chance of incorporating errors in the DNA. The diversity of HIV populations in a single patient originates from the combination of generating genetic mistakes and the high viral production rate. In addition, the high mutation rate also results in a rapid resistance onset against antiretroviral treatment.

Upon entering the host cell nucleus, the integrase enzyme integrates the viral DNA into the human genome. After this irreversible process, the integrated viral genetic material, designated as provirus, can persist in the host genome for

many days, even years, acting as a latent viral reservoir. The viral genome can be activated at a later time to produce new virus particles. This process is initiated by transcription of the genome into RNA transcripts and successive transport of these strands to the cytoplasm, where the genes are translated into proteins. In the early phase, only small regulatory proteins are expressed, such as Tat, Rev and Nef. Tat stimulates the provirus transcription process by RNA polymerase II. In the late phase, multiple Rev molecules have been accumulated and will collaborate to export the late mRNA transcripts to the cytoplasm by binding to the Rev responsive element (RRE) on unspliced and partially spliced mRNA strands. These unspliced and partially spliced mRNAs encode for the viral structural and enzymatic proteins, which assemble at the plasma membrane to form new particles. Budding of these particles ultimately results in expulsion of new HIV virions. During the maturation step, protease enzymes process multiple viral proteins, converting these into infective HIV virions. Those new virus particles can subsequently infect other host cells and spread the HIV infection over the entire human body.

## 1.4 HIV gp41

During the fusion step, the gp41 protein is responsible for connecting the host and viral membrane. The protein consists of three distinct domains: an internal domain called the endodomain, a transmembrane domain and an external domain called the ectodomain [16, 17] (see Figure 1.3A). The ectodomain can also be divided in different structural regions, the N-terminal fusion peptide (FP), the fusion peptide proximal region (FPPR), two heptad repeats (HR) called the N-terminal HR (N-HR) and C-terminal HR (C-HR) and a membrane proximal external region (MPER).

The gp41 protein undergoes major conformational changes during the fusion process initiated by receptor and coreceptor binding to gp120 [14, 15, 19–21]. A scheme of the fusion process is shown in figure 1.3C. After insertion of the FP in the cellular membrane, a pre-fusogenic state is formed where the N- and C-terminal heptad repeats adopt an extended trimeric coiled coil connecting both viral and cellular membrane (figure 1.3B). This pre-hairpin intermediate is subsequently converted into a stable antiparallel six-helix bundle arrangement by folding of the three C-terminal HRs onto the grooves of the trimeric N-terminal HR bundle. As a result, the fusion peptide and transmembrane domain are now oriented in the same direction, thereby bringing viral and cellular membranes in close proximity. Finally, a fusogenic state is formed where both membranes are fused resulting in a postfusion conformation, which is thought to be stabilised by the membrane anchoring regions FPPR and MPER [22, 23].

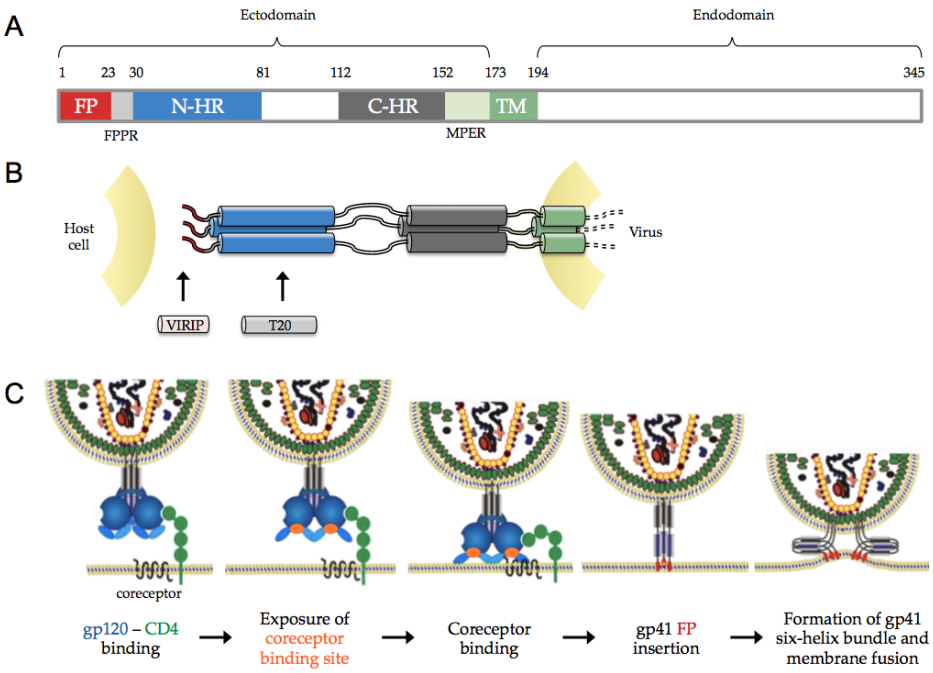


Figure 1.3: **Overview of the gp41 structure.** (A) Sequence of gp41. The protein is divided into an ecto- and endodomain. The ectodomain comprises a N-terminal fusion peptide (FP), a fusion peptide proximal region (FPPR), a N-terminal and C-terminal heptad repeat (N-HR and C-HR, respectively) and a membrane proximal external region (MPER) close to the viral transmembrane domain (TM). (B) Schematic representation of the gp41 structure in the pre-hairpin intermediate step. During this step, the structural trimeric bundles (N-HR and C-HR) adopt an extended conformation, which allows insertion of the N-terminal FP in the host cell membrane. Membrane fusion can be inhibited by preventing the FP insertion event (by VIRIP and its derivatives) or by competitive binding to N-HR (by T20), which prevent collapse of the trimeric coil in a six-helix bundle structure. (C) Illustration of the HIV membrane fusion. First, binding of gp120 to the CD4 receptor triggers a conformational change, revealing the coreceptor binding site. After binding of the coreceptor, the HIV virion lies close to the host cell membrane. Next, gp41 shoots its N-terminal FP into the target cell membrane with a hairpin-like mechanism (gp120 and (co-)receptors are omitted here for clarity). Formation of a gp41 six-helix bundle subsequently provides energy to fuse viral and cellular membranes. Figure adapted from Forssmann *et al.* [18]

Up to now, the entire structure of gp41 remains unsolved. However, subdomains have been crystallised, including HR peptide mixtures, fused constructs and inhibitor bound forms. Molecular dynamics (MD) simulations have been employed by different groups to study the conformational changes in the gp41 protein and to probe the interactions of potential inhibitors [24–28].[1]

[1] We have contributed to this field by exploring the conformational ensemble of gp41 FP in solution, as discussed in chapter 5.

### 1.4.1 Entry inhibitors

Targeting the entry process of HIV-1 is a promising approach, which has already led to the development of two clinically approved antiretroviral drugs, that is, T20 (also known as enfuvirtide, marketed as Fuzeon by Roche/Trimeris) [29, 30] and maraviroc (marketed as Selzentry/Celsentri by Pfizer) [31].[2] Maraviroc was developed by medicinal chemistry optimisation of a hit compound identified from high-troughput screening experiments. The compound inhibits viral entry indirectly by blocking attachment of the viral particle to the cellular coreceptor CCR5. In contrast, T20 is a peptidic inhibitor that mimics the C-terminal HR residues 127 to 162 by binding to the N-terminal HR of gp41, thereby inhibiting the six-helix bundle formation essential for fusion. Retroviral inhibitory peptides were identified serendipitously in a vaccine development experiment at Duke university. T20 is an optimised form, developed in collaboration with the pharmaceutical industry. Unfortunately, a number of reports have shown that the efficacy of T20 diminishes in long-term clinical studies due to the emergence of resistant viral strains [32–37]. Resistance mutations have been shown to manifest in a 10 amino acid motif in the N-terminal HR, while specific mutations in C-terminal HR are able to restore the fusion process [38].

To explore the origins of resistance against T20, McGillick *et al.* [24] used an extensive MD simulation and binding free energy calculation study of a docked gp41-T20 complex embedded in an explicit lipid membrane. Using an improved gp41 model based on a SIV gp41 ectodomain structure [39, 40], a good quantitative agreement with experimental resistance data was found. The structural information from the MD simulation study has been used to design novel indole compounds [41] and in later resistance studies as well [42]. Furthermore, they also found a significant number of favourable interactions between T20 and the lipid bilayer that stabilise the membrane-protein complex. This notion suggests that peptides interacting favourably with both gp41 and the membrane can have an increased inhibitor potency. Interestingly, a peptide derived from the C-terminal HR containing an attached cholesterol group displayed increased antiviral efficiency, confirming that such an approach is indeed feasible [43].

A study by Singh *et al.* [44] demonstrated that extending the N- and C-terminal ends of T20 to a total of 42 residues increases helicity of the peptide in solution. In a series of MD simulation articles, Martins do Canto *et al.* [45–49] studied the interactions of T20 and a second-generation fusion inhibitor T1249 alone in solution and in the presence of modelled membranes. In the studied time scale, they found that the peptides adopt a mainly disordered structure in solution [45],

[2] Structures of T20 and maraviroc can be found in reference [5].

while a mainly $\pi$-helical structure was found in the presence of membranes [46–49]. In addition, fusion inhibitors seem to bind less deeply to the liquid-ordered membrane containing cholesterol and T1249 interacts stronger with the membrane compared to T20, as shown previously by fluorescence experiments [50].

In addition to T20, a notable number of other C-terminal HR derived peptides have been investigated. For example, C34 and derived mutants (corresponding to the N-terminal HR residues of 117 to 150) have been studied in complex with an N-terminal HR trimer using a combined MD and binding free energy analysis approach [28]. A good correlation with experimental results was found despite the observed experimental small binding affinity range. In addition, residue decomposition of the binding free energy revealed the hot spot residues contributing to the interaction. Using this approach, the importance of a conserved hydrophobic pocket of gp41 as an attractive drug target [17] was validated by the presence of strong van der Waals energy values between gp41 residues and the C34 peptide. Another MD study was performed by Hartono *et al.* [51] on a small peptide derived from human lysozyme. This peptide called HL9 contains only nine residues and could potentially bind through two Trp residues into the hydrophobic pocket of the gp41 ectodomain. Multiple docking poses were explored and calculated binding free energy values were in fair agreement with experimental measurements.

### 1.4.2 The anchoring inhibitor VIRIP

Most entry inhibitor peptides and small molecules target the hydrophobic binding pocket of gp41. Unfortunately, virus resistance against T20 and the latest generation of fusion inhibitors is caused by amino acid substitutions in this region, thus leading to viral escape from entry inhibitor therapy. While T20 resistance is mainly caused by single amino acid substitutions in the binding pocket [33], resistance against newer entry inhibitors such as T1249 and T2635 requires multiple mutations within gp41 [35]. Hence, it is paramount and appealing to target parts of gp41 that are highly conserved. The fusion peptide (FP) is one of the most conserved parts of gp41 and recently a 23 amino acid peptide has been found to show inhibitory activity against this part [18, 52]. This peptide called VIRIP (VIRus INhibitory Peptide) is naturally occurring in human blood and was identified by screening of a comprehensive blood-derived peptide/protein library.[3] VIRIP inhibits HIV-1 entry by binding to the gp41 FP, thus preventing its insertion into the target cell membrane and successive membrane fusion. These results were

[3]Unfortunately, the wild type sequence of VIRIP is not potent enough to suppress viral levels in blood.

unexpected, since gp41 FP is only temporarily exposed during the fusion process [14]. Based on the wild type sequence of VIRIP (LEAIPMSIPPEVKFNKPFVF), a number of derivatives have been developed such as VIR-165 (LEAIPCSIPPC-FAFNKPFVF) and a dimeric form, VIR-576 (LEAIPCSIPPEFLFGKPFVFx2). VIRIP derivatives are promising for clinical development because of their low toxicity and immunogenicity, broad activity against HIV-1, and lack of cross-resistance with other drugs, including those with resistance against T20. An extensive structure-activity relationship study allowed the design of VIRIP derivatives with increased affinity and antiviral potency [52]. Recently, one of these derivatives (VIR-576) was evaluated in a clinical phase I/II trial in treatment of naive HIV-1 infected individuals [18]. The results showed that VIR-576 monotherapy was well tolerated and reduced the plasma viral loads by more than one order of magnitude [18]. Long-term experiments illustrated that more than 450 days were required to generate resistance against optimised VIRIP derivates [53, 54]. Moreover, the resistant mutations significantly decreased the replicative capacity of the HIV strains [54]. To improve the affinity against gp41 FP even further, a detailed understanding of the structure of the FP fragment is needed. Unfortunately, experimental structures of the FP have been limited to membrane [55–58] or inhibitor-bound forms [52], so no free form of the gp41 FP is known. Interestingly, other inhibitors have also been found to interact with FP, such as synthetic hexapeptides [59] and peptides of the E2 envelope protein of Hepatitis G virus (HGV) [60–66].

### 1.4.3 The gp41 FP structure

As mentioned previously, the gp41 FP is not only an interesting target for antiretroviral therapy, it also plays a crucial role during the fusion process by constituting the first viral contact with the cellular membrane. Not surprisingly, a lot of research has been dedicated to elucidate the secondary structure of the FP in the membrane. The FP sequence (AVGIGALFLGFLGAASTMGARS) is highly hydrophobic [67] and mutational analysis revealed critical glycine residues in the FP sequence to fuse viral and cellular membranes [68, 69] and unfavourable polar substitutions [70–72]. Nevertheless, until today the secondary structure of the FP remains under discussion, due to conflicting experimental reports of $\alpha$-helical [55–58, 73], $\beta$-sheet [74–82] and disordered conformations [83, 84], or a combination of these [85–98]. An overview of the experimentally determined structures is found in Figure 1.4 and Table 1.1. Analysis is blurred due to use of different buffer conditions, peptide concentrations, or FP constructs with different lengths and solubility. In addition, possible oligomerisation and subsequent aggregation issues, absence of single-molecule studies and influence of lipid composition and lipid constructs (e.g. micelles versus lipid bilayers) prevent a clear interpretation. Consequently, it is

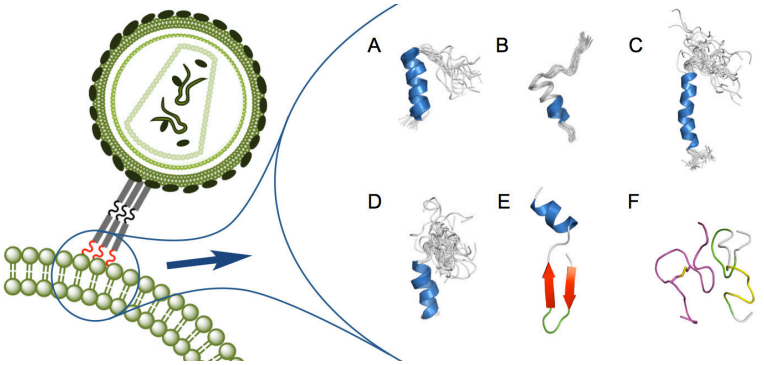


Figure 1.4: **Schematic representation of HIV-1 fusion and overview of the experimentally determined structures of gp41 FP in membrane-like environments or solution.** After recognition of the host cell membrane CD4 receptor and co-receptors by the gp120 protein (omitted for clarity), the trimeric gp41 protein (grey) undergoes a conformational rearrangement, thereby exposing and inserting the gp41 FP (red) in the host cell membrane. The inlet on the right displays some of the proposed structures of gp41 FP (A-F). (A) $\alpha$-helical FP23 with FTIR restraints in hexafluoroisopropanol (PDB-code: 1ERF) [55]. (B) $\alpha$-helical FP23 with FTR restraints in SDS micelles (PDB-code: 1P5A) [56]. (C) Solution NMR $\alpha$-helical FP30 (PDB-code: 2ARI) [57]. (D) Solution NMR $\alpha$-helical FP23 (PDB-code: 2PJV) [58]. (E) $\beta$-sheet model FP23 (model based on PDB-code: 3D58) [24]. (F) Solution NMR unstructured FP23 in complex with VIR-165 (pink), a disulphide stabilised derivative of the fusion peptide inhibitor VIRIP (PDB-code: 2JNR) [52]. FP secondary structures are coloured according to the DSSP convention with blue: $\alpha$-helix, red: $\beta$-sheet, white: coil, black: $\beta$-bridge, green: bend, yellow: turn, purple: $\pi$-helix and grey: 3-10-helix. Figure taken from Venken *et al.* [110].

questioned if the peptide surroundings are physiologically relevant to understand membrane fusion in all the previously mentioned studies. Even experiments under identical sample conditions showed the presence of both $\alpha$-helical and $\beta$-sheet structures at the same time [91, 97, 99]. Thus, it has been put forward that not just the peptide length or structure determines fusogenicity, but also other factors such as the membrane dipole potential [92, 100], insertion depth [81, 94, 97], insertion angle [85, 88, 89, 101–106], membrane curvature [90, 107] and corresponding membrane kinetics [91, 108]. It has also been suggested that the cholesterol content in lipid bilayers could play an important role in mediating membrane curvature [91–95, 97, 109].

Because many experimental techniques investigate only a time-averaged ensemble of conformations, atomistic MD simulations are an alternative approach to probe the interactions of gp41 FP with model membrane systems. A monomeric membrane-bound structure of gp41 FP can be representative for the initial moment of membrane insertion. One of the first studies was performed by Kamath and Wong [25, 111] who studied an $\alpha$-helical conformation of different mutants and the wild type form of gp41 FP (first 16 residues) in an equilibrated lipid bilayer for 1.4 to 2 ns. Contrary to the mutant forms that adopt conformations parallel to the

Table 1.1: **Overview of gp41 FP structural data**

| PDB ID | Sequence | Method | # Models | Year | Authors |
|---|---|---|---|---|---|
| 1ERF | AVGIGALFLGFLGAAGSTMGARS-NH2 | FTIR | 17 | 2000 | Gordon *et al.* [55] |
| 1P5A | AVGIGALFLGFLGAAGSTMGARS-NH2 | FTIR | 19 | 2003 | Gordon *et al.* [56] |
| 2ARI | AVGIGALFLGFLGAAGSTMGAASMTLTVQA | NMR | 30 | 2005 | Jaroniec *et al.* [57] |
| 2JPV | AVGIGALFLGFLGAAGSTVGAASG | NMR | 38 | 2007 | Li *et al.* [58] |
| 2JNR[a] | AVGIGALFLGFLGAAGSTMGARS | NMR | 1 | 2007 | Münch *et al.* [52] |

(a) This FP structure was obtained in complex with VIR-165, a VIRIP derivative with sequence LEAIPCSIPPC-FAFNKPFVF.

lipid bilayer, the wild type adopts an obliquely inserted angle in the membrane. In addition, the results indicate that fusogenicity is not explained by variations in the FP secondary structure, but are rather caused by dissimilar conformational flexibility patterns between wild type FP and inactive mutants. While these conclusions are in line with experimental measurements, a V2E mutant inserts deeper into membranes than wild type FP in experiments [112]. Barz *et al.* [90] conducted an extensive MD study by simulating both $\alpha$-helical and $\beta$-sheet FP dimers. To reduce bias of the starting conformation of the modelled system, the initial $\alpha$-helical conformations were modelled both parallel and perpendicular to the bilayer normal. In addition, $\alpha$-helix to $\beta$-sheet transitions were performed by resorting to steered MD (SMD) simulations [90]. While the FP adopts a broad variety of conformations in solution, the diffusion time is much slower in membrane environments, hence non-equilibrium simulations were required to speed up the secondary structure transition process. This requirement was demonstrated recently by 100 ns long simulations of FPs with different length (both 17 and 23 residues) and charge of termini, which either remained $\alpha$-helical or partially unfolded during the simulation trajectory [113]. In summary, Barz *et al.* [90] compared their simulation data to Fourier transform infrared spectroscopy (FTIR) measurements and suggested that the FP structure depends on the area per lipid of the membrane surface. In contrast, Grasnick *et al.* [84] postulated that not an ordered conformation but rather an irregular FP structure determines fusogenicity, based on MD simulations using nuclear magnetic resonance (NMR) orientational constraints [84]. Although a large number of possible structures can fulfil the experimentally measured constraints, the found peptide conformations were not compatible with any regular secondary structure. Thus, while the experimental methods resulted in many different suggestions for the role of the gp41 FP during membrane fusion, the modelling studies have spurred contradicting results as well. As such, despite a growing amount of evidence of balanced interplay between peptide and membrane interactions, the crucial role of the gp41 FP remains to be investigated more in detail in the future.

## 1.5 The HIV Rev protein

Transcription of new HIV RNA strands begins in the nucleus after integration of the viral DNA in the host cell genome. During the early stage, splicing of the viral RNA generates fully spliced RNA transcripts. These RNA strands are actively transported to the cytoplasm by the general cellular RNA export pathway and encode for the viral regulatory proteins. In the late phase of the viral life cycle, larger unspliced (intron-containing) viral RNA transcripts are expressed. The transport of these larger RNA transcripts depends on more sophisticated transport mechanisms. This transport process requires a delicate interplay between viral and cellular proteins and is mediated by the Rev protein (Regulator of Expression of Virion proteins) [114, 115]. This viral protein is indispensable for the onset of the late replication cycle of HIV. In fact, Rev is one of those small proteins that are translated first, next to other regulatory proteins such as Tat and Nef. Longer intron-containing RNA strands code for larger essential proteins such as Gag, Pol and Env. Nuclear export of those strands is initiated once a substantial concentration of Rev protein is present in the nucleus. Viral replication is abrogated without the synthesis of late viral proteins and inhibition of the biological function Rev would therefore be an attractive novel antiretroviral strategy.

Although Rev mainly resides in the nucleus, it has been shown to continuously shuttle between the nucleus and the cytoplasm by exploiting the cellular CRM1-mediated export machinery [116]. In that way, it transports the viral RNA to the cytoplasm, allows translation of new viral proteins and is subsequently recycled back into the nucleus. As such, the Rev protein is not only important for the HIV infection progress, but it also constitutes a paradigm for the study of nuclear export mechanisms in cells.

### 1.5.1 The Rev domain organisation

Rev is a protein of approximately 19 kD located predominantly in the host cell nucleus. It consists of 116 amino acids decomposed in distinct domains, which are visualised in Figure 1.5. The N-terminal region contains a nuclear entry inhibitory signal (NIS) of 15 amino acids important for nucleo-cytoplasmatic trafficking, and is thought to adopt an amphipathic helical structure [117]. In addition, an arginine-rich motif (ARM) of residues 34-50 serves a dual role. First, it functions as a nuclear localisation signal (NLS). This motif binds to the host protein importin-$\beta$ [118] and returns the Rev protein to the nucleus. Second, the region acts as an RNA binding domain by attachment to the Rev response element (RRE) on RNA, a stem loop of 351 nucleotides located in the *env* gene of unspliced RNA strands [119].

The interaction between Rev and the RRE is specific as only a few mutations in Rev [120] or in the RRE [121] can be sufficient to reduce the binding. The interaction is thought to be exclusively entropically driven [122] owing to shape recognition of the flexible ARM arginine side chains with the RRE with only a limited amount of base-specific hydrogen bonds [123, 124]. Furthermore, it has been suggested that the ARM region can be disordered but may fold upon binding of the RRE [125].

A third essential motif is the nuclear export signal (NES), which resides in the intrinsically unstructured C-terminal domain [126]. That leucine-rich motif interacts with cellular proteins such as CRM1 (also known as exportin1) [127]. However, there is a marked difference in key hydrophobic residue spacing between cellular classic NES and viral NES. To effectively bind CRM1, the retroviral NES adopts an extended conformation, while the NES of cellular proteins such as protein kinase inhibitor (PKI) folds as an $\alpha$-helix [128].

In sum, the NLS and NES motifs are essential for Rev-shuttling between the nucleus and the cytoplasm; while NLS is dedicated to nuclear import guided by importin-$\beta$, in contrast NES is required for export to the cytoplasm by CRM1. By using this mechanism, HIV has found a way to bypass the nuclear retention of its intron-containing RNA by linking these RNA species to a cellular protein transport mechanism (CRM1, importin-$\beta$ and several other host cell co-factors). In addition to CRM1, many other cellular proteins like the DEAD box helicases have been identified as cofactor for Rev function. It has been suggested that these helicases, such as DDX1 [129–131], DDX3 [132] and DDX5 [133], can synergistically modulate the function of Rev in the host cell [134]. Hence, the Rev protein requires a considerable amount of conformational flexibility to attain a combination of cooperative protein-RNA and protein-protein interactions.

In addition to the previously mentioned interactions, another degree of complexity arises from the multimerisation of the Rev protein. It has been found that multiple Rev monomers can bind to a single RRE forming a functional multimeric complex [135–137]. However, Rev is also able to multimerise in the absence of RNA [138]. The Rev multimerisation domains can be found in the N-terminal regions overlapping with the NIS site (M1) and in the C-terminal regions upstream from the NLS domain (M2).

### 1.5.2 The Rev structure

Previously, little was known about the three-dimensional structure of the HIV Rev multimer. A wide array of biophysical studies have been pursued to decipher the nature of the Rev protein function, for example: mutational analysis [140–

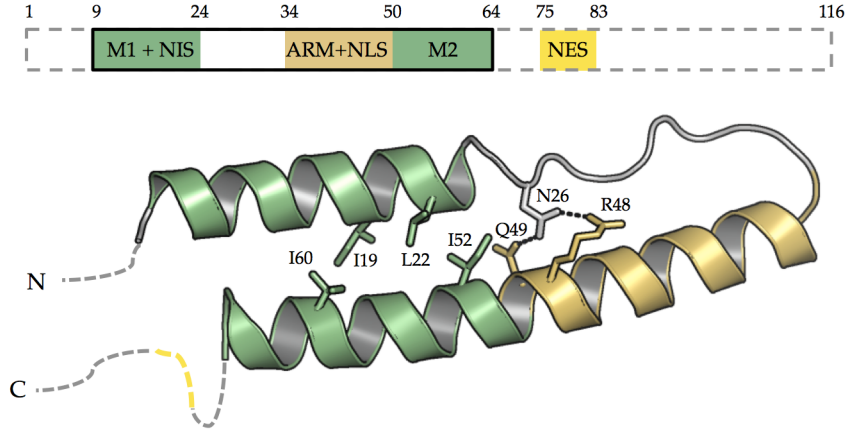


Figure 1.5: **Overview of the Rev domain structure.** Crystallised parts are highlighted with black boxes. Unresolved regions are visualised with grey stripes. The Rev structure is divided in two multimerisation domains (M1 + M2). The first multimerisation domain overlaps with a nuclear entry inhibitory signal (NIS). Similarly, the arginine rich motif (ARM) functions as a nuclear localisation signal (NLS). Finally, a nuclear export signal (NES) can be found in the intrinsically disordered C-terminal region. The Rev helix-loop-helix motif is stabilised by hydrophobic residues in the multimerisation domains (residues I19, L22, I52 and I60) and by a hydrogen bond network in the ARM (residues N26, R48 and Q49). Image generated from PDB structure 2X7L [139].

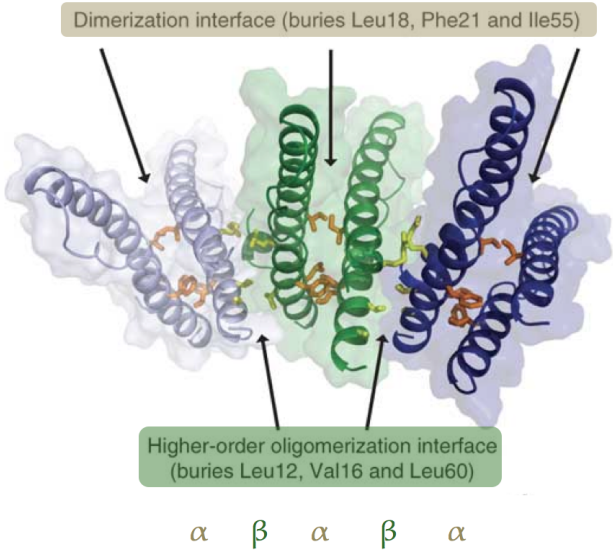


Figure 1.6: **Rev multimerisation.** A Rev multimer can contain multiple interfaces, a dimerisation interface $\alpha$ comprising residues Leu18, Phe21 and Ile55, and a higher-order multimerisation interface $\beta$ composed of residues Leu12, Val16 and Leu60. These interfaces are repeated in a symmetrical pattern in a hexamer as $\alpha$, $\beta$, $\alpha$, $\beta$, $\alpha$. Image adapted from Daugherty *et al.* [140].

142], circular dichroism (CD) spectroscopy [143, 144], single-molecule fluorescence spectroscopy [145], Förster resonance energy transfer (FRET) measurements [146], atomic force microscopy (AFM) [147, 148], and solid state NMR [149]. Those experiments suggest that the Rev monomer consists of a helix-turn-helix motif. It was hypothesised that multiple Rev molecules cooperate to form a V-shaped protein that clamps the viral RNA at the RRE motif. The Rev monomers self-associate one by one after binding to the RRE to form a symmetrical interface pattern: $\alpha$, $\beta$, $\alpha$, $\beta$, ... as illustrated in figure 1.6. Thus, two different symmetrical interfaces are present in the Rev protein [139–141]. We applied the following naming convention throughout this thesis: a dimerisation interface $\alpha$, which is the binding interface occurring first upon formation of a multimeric Rev protein-protein complex, and a higher order oligomerisation interface $\beta$.

Recently, a long-awaited structure of the Rev protein was "Rev-ealed" (pun intended by Hammarskjold *et al.* [150]) by two independent groups [139, 140]. One crystal structure consists of the wild type dimeric structure of Rev in complex with a monoclonal Fab fragment (PDB entry: 2X7L), comprising the Rev $\beta$ interface but lacking the $\alpha$ interface [139]. A second crystal structure consists of a mutant Rev complex (L12S and L60R) containing both $\alpha$ and $\beta$ interfaces (PDB entry: 3LPH) [140]. Although the Rev structure is only partially resolved (the flexible C-terminal part is intrinsically unfolded), the helix-turn-helix motif containing the sites of multimerisation is present in both solved crystal structures and provides insight into the molecular interactions thriving multimer formation. Backbone superimposition of the monomer conformations from both crystals results in a significant structural similarity (root mean square deviation (RMSD) < 1 Å). Furthermore, an angle of 120° was reported between the $\beta$ interface monomers of 3LPH, while a broader angle of 140° is present between the 2X7L monomers. An overview of the released crystal structures can be found in Table 1.2.

Table 1.2: **Overview of crystal structure data of Rev**

| PDB ID | Sequence | Residues | Structure | Resolution | R-value | R-free | Year | Authors |
|---|---|---|---|---|---|---|---|---|
| 2X7L[a] | wild type | 9 to 65 | dimer ($\beta$) | 3.17 Å | 0.235 | 0.250 | 2010 | Dimattia *et al.* [139] |
| 3LPH[b] | L12S & R60L | 8 to 70 | tetramer ($\alpha$-$\beta$-$\alpha$) | 2.5 Å | 0.228 | 0.261 | 2010 | Daugherty *et al.* [140] |

(a) The 2X7L structure also contains two specifically engineered monoclonal Fab fragments that stabilise each monomer of the dimeric complex. (b) Not all residues are resolved in each monomer.

### 1.5.3 Multimerisation

The multimerisation of Rev molecules occurs at flanking sequences of the arginine rich motif (ARM), which functions both as a nuclear localisation signal (NLS) and as a RRE-binding site. The multimerisation process is essential for the function of Rev [137, 151]. In fact, a single Rev molecule bound to the RRE of spliced mRNA is not able to promote nuclear export [151, 152]. Mutant Rev proteins are still able to bind to the RRE and form dimers, but the mutations disrupt higher order oligomerisation and render Rev export deficient [141, 143]. A number of multimerisation hot spot residues were identified and it was suggested that L18 and I55 are essential for the $\alpha$ interface while L12, V16 and L60 are important for the $\beta$ interface.

Currently known inhibitors target the Rev/RRE [153–155] and Rev/CRM1 interactions [156–158]. Unfortunately, the first easily induce resistance due to compensatory nucleotide changes in the RRE [159], while the latter approach can also disturb nucleocytoplasmatic trafficking of the host. The disruption of the multimerisation process, which only targets viral proteins, could therefore be an alternative approach to inhibit propagation of HIV-1 inside the host cells. Recently, a llama single-domain antibody designated $Nb_{190}$ has been discovered that blocks the multimerisation site of Rev and thereby disrupts multimerisation of Rev *in vitro* and *in vivo* [160–162]. This nanobody interacts with residues Lys20 and Tyr23 in the multimerisation domain of the Rev protein. Therefore, the multimerisation process is considered an interesting novel target for the inhibition of HIV replication.

## References

[1] Barré-Sinoussi, F., Chermann, J. C., Rey, F., Nugeyre, M. T., Chamaret, S., Gruest, J., Dauguet, C., Axler-Blin, C., Vézinet-Brun, F., Rouzioux, C., Rozenbaum, W., and Montagnier, L. (1983). Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science*. 220(4599):868–871.

[2] Gallo, R. C., Salahuddin, S. Z., Popovic, M., Shearer, G. M., Kaplan, M., Haynes, B. F., Palker, T. J., Redfield, R., Oleske, J., and Safai, B. (1984). Frequent detection and isolation of cytopathic retroviruses (HTLV-III) from patients with AIDS and at risk for AIDS. *Science*. 224(4648):500–503.

[3] Silvestri, G. (2013). Embracing the complexity of HIV immunology. *Immunol. Rev*. 254(1):5–9.

[4] Mitsuya, H., Weinhold, K. J., Furman, P. A., St Clair, M. H., Lehrman, S. N., Gallo, R. C., Bolognesi, D., Barry, D. W., and Broder, S. (1985). 3'-Azido-3'-deoxythymidine (BW A509U): an antiviral agent that inhibits the infectivity and cytopathic effect of human T-lymphotropic virus type III/lymphadenopathy-associated virus in vitro. *Proc. Natl. Acad. Sci. U. S. A*. 82(20):7096–7100.

[5] De Clercq, E. (2009). Anti-HIV drugs: 25 compounds approved within 25 years after the discovery of HIV. *Int. J. Antimicrob. Ag*. 33(4):307–320.

[6] Katlama, C., Deeks, S. G., Autran, B., Martinez-Picado, J., van Lunzen, J., Rouzioux, C., Miller, M., Vella, S., Schmitz, J. E., and Ahlers, J. (2013). Barriers to a cure for HIV: new ways to target and eradicate HIV-1 reservoirs. *Lancet*. 381(9883):2109–2117.

[7] Hütter, G., Nowak, D., Mossner, M., Ganepola, S., Müssig, A., Allers, K., Schneider, T., Hofmann, J., Kücherer, C., Blau, O., Blau, I. W., Hofmann, W. K., and Thiel, E. (2009). Long-term control of HIV by CCR5 Delta32/Delta32 stem-cell transplantation. *N. Engl. J. Med*. 360(7):692–698.

[8] Allers, K., Hutter, G., Hofmann, J., Loddenkemper, C., Rieger, K., Thiel, E., and Schneider, T. (2011). Evidence for the cure of HIV infection by CCR5 32/ 32 stem cell transplantation. *Blood*. 117(10):2791–2799.

[9] Henrich, T. J. and Kuritzkes, D. R. (2012). HIV-1 entry inhibitors: recent development and clinical use. *Curr. Opin. Virol*. 1–7.

[10] Kent, S. J., Reece, J. C., Petravic, J., Martyushev, A., Kramski, M., De Rose, R., Cooper, D. A., Kelleher, A. D., Emery, S., Cameron, P. U., Lewin, S. R., and Davenport, M. P. (2013). The search for an HIV cure: tackling latent infection. *Lancet Infect. Dis*. 13(7):614–621.

[11] World Health Organization and UNAIDS (2011). World Health Organization: Global HIV/AIDS response. *Geneva: World Health Organization*.

[12] Voet, A. (2010). Development of small molecule protein-protein interaction inhibitors targeting the HIV-1 integrase-LEDGF/p75 complex. *PhD thesis*.

[13] Frankel, A. D. and Young, J. A. (1998). HIV-1: fifteen proteins and an RNA. *Annu. Rev. Biochem*. 67(1):1–25.

[14] Doms, R. W. and Moore, J. P. (2000). HIV-1 membrane fusion: targets of opportunity. *J. Cell Biol*. 151(2):F9–14.

[15] Eckert, D. M. and Kim, P. S. (2001). Mechanisms of viral membrane fusion and its inhibition. *Annu. Rev. Biochem*. 70:777–810.

[16] Chan, D. C., Fass, D., Berger, J. M., and Kim, P. S. (1997). Core structure of gp41 from the HIV envelope glycoprotein. *Cell*. 89(2):263–273.

[17] Chan, D. C., Chutkowski, C. T., and Kim, P. S. (1998). Evidence that a prominent cavity in the coiled coil of HIV type 1 gp41 is an attractive drug target. *Proc. Natl. Acad. Sci. U. S. A*. 95(26):15613–15617.

[18] Forssmann, W. G., The, Y. H., Stoll, M., Adermann, K., Albrecht, U., Tillmann, H. C., Barlos, K., Busmann, A., Canales-Mayordomo, A., Gimenez-Gallego, G., Hirsch, J., Jimenez-Barbero, J., Meyer-Olson, D., Munch, J., Perez-Castells, J., Standker, L., Kirchhoff, F., and Schmidt, R. E. (2010). Short-Term Monotherapy in HIV-Infected Patients with a Virus Entry Inhibitor Against the gp41 Fusion Peptide. *Sci. Transl. Med*. 2(63):63re3.

[19] Epand, R. M. (2003). Fusion peptides and the mechanism of viral fusion. *Biochim. Biophys. Acta*. 1614(1):116–121.

[20] Gallo, S. A., Finnegan, C. M., Viard, M., Raviv, Y., Dimitrov, A., Rawat, S. S., Puri, A., Durell, S., and Blumenthal, R. (2003). The HIV Env-mediated fusion reaction. *Biochim. Biophys. Acta*. 1614(1):36–50.

[21] Harrison, S. C. (2008). Viral membrane fusion. *Nat. Struct. Mol. Biol*. 15(7):690–698.

[22] Buzón, V., Natrajan, G., Schibli, D., Campelo, F., Kozlov, M. M., and Weissenhorn, W. (2010). Crystal Structure of HIV-1 gp41 Including Both Fusion Peptide and Membrane Proximal External Regions. *PLoS Pathog*. 6(5):e1000880.

[23] Lakomek, N.-A., Kaufman, J. D., Stahl, S. J., Louis, J. M., Grishaev, A., Wingfield, P. T., and Bax, A. (2013). Internal Dynamics of the Homotrimeric HIV-1 Viral Coat Protein gp41 on Multiple Time Scales. *Angew. Chem. Int. Ed*. 52(14):3911–3915.

[24] McGillick, B. E., Balius, T. E., Mukherjee, S., and Rizzo, R. C. (2010). Origins of Resistance to the HIVgp41 Viral Entry Inhibitor T20. *Biochemistry*. 49(17):3575–3592.

[25] Kamath, S. and Wong, T. C. (2002). Membrane structure of the human immunodeficiency virus gp41 fusion domain by molecular dynamics simulation. *Biophys. J*. 83(1):135–143.

[26] Kim, J. H., Hartley, T. L., Curran, A. R., and Engelman, D. M. (2009). Molecular dynamics studies of the transmembrane domain of gp41 from HIV-1. *Biochim. Biophys. Acta*. 1788(9):1804–1812.

[27] Tan, J. J., Chen, W. Z., and Wang, C. X. (2006). Investigating interactions between HIV-1 gp41 and inhibitors by molecular dynamics simulation and MM–PBSA/GBSA calculations. *J. Mol. Struc. (Theochem.)*. 766(2-3):77–82.

[28] Strockbine, B. and Rizzo, R. C. (2007). Binding of antifusion peptides with HIVgp41 from molecular dynamics simulations: Quantitative correlation with experiment. *Proteins*. 67(3):630–642.

[29] Kilby, J. M., Hopkins, S., Venetta, T. M., DiMassimo, B., Cloud, G. A., Lee, J. Y., Alldredge, L., Hunter, E., Lambert, D., Bolognesi, D., Matthews, T., Johnson, M. R., Nowak, M. A., Shaw, G. M., and Saag, M. S. (1998). Potent suppression of HIV-1 replication in humans by T-20, a peptide inhibitor of gp41-mediated virus entry. *Nat. Med*. 4(11):1302–1307.

[30] Matthews, T., Salgo, M., Greenberg, M., Chung, J., DeMasi, R., and Bolognesi, D. (2004). Case history: Enfuvirtide: the first therapy to inhibit the entry of HIV-1 into host CD4 lymphocytes. *Nat. Res. Drug. Discov*. 3(3):215–225.

[31] Fätkenheuer, G., Pozniak, A. L., Johnson, M. A., Plettenberg, A., Staszewski, S., Hoepelman, A. I. M., Saag, M. S., Goebel, F. D., Rockstroh, J. K., Dezube, B. J., Jenkins, T. M., Medhurst, C., Sullivan, J. F., Ridgway, C., Abel, S., James, I. T., Youle, M., and van der Ryst, E. (2005). Efficacy of short-term monotherapy with maraviroc, a new CCR5 antagonist, in patients infected with HIV-1. *Nat. Med*. 11(11):1170–1172.

[32] Wei, X., Decker, J. M., Liu, H., Zhang, Z., Arani, R. B., Kilby, J. M., Saag, M. S., Wu, X., Shaw, G. M., and Kappes, J. C. (2002). Emergence of Resistant Human Immunodeficiency Virus Type 1 in Patients Receiving Fusion Inhibitor (T-20) Monotherapy. *Antimicrob. Agents Chemother*. 46(6):1896–1905.

[33] Greenberg, M. L. (2004). Resistance to enfuvirtide, the first HIV fusion inhibitor. *J. Antimicrob. Chemother*. 54(2):333–340.

[34] Baldwin, C. E., Sanders, R. W., Deng, Y., Jurriaans, S., Lange, J. M., Lu, M., and Berkhout, B. (2004). Emergence of a Drug-Dependent Human Immunodeficiency Virus Type 1 Variant during Therapy with the T20 Fusion Inhibitor. *J. Virol*. 78(22):12428–12437.

[35] Eggink, D., Langedijk, J. P. M., Bonvin, A. M. J. J., Deng, Y., Lu, M., Berkhout, B., and Sanders, R. W. (2009). Detailed Mechanistic Insights into HIV-1 Sensitivity to Three Generations of Fusion Inhibitors. *J. Biol. Chem*. 284(39):26941–26950.

[36] Miller, M. D. and Hazuda, D. J. (2004). HIV resistance to the fusion inhibitor enfuvirtide: mechanisms and clinical implications. *Drug Resist. Updates*. 7(2):89–95.

[37] Berkhout, B., Eggink, D., and Sanders, R. W. (2012). Is there a future for antiviral fusion inhibitors? *Curr. Opin. Virol*. 2(1):50–59.

[38] Ray, N., Blackburn, L. A., and Doms, R. W. (2009). HR-2 Mutations in Human Immunodeficiency Virus Type 1 gp41 Restore Fusion Kinetics Delayed by HR-1 Mutations That Cause Clinical Resistance to Enfuvirtide. *J. Virol*. 83(7):2989–2995.

[39] Caffrey, M., Cai, M., Kaufman, J., Stahl, S., Wingfield, P., Covell, D., Gronenborn, A., and Clore, G. (1998). Three-dimensional solution structure of the 44 kDa ectodomain of SIV gp41. *Embo Journal*. 17(16):4572–4584.

[40] Caffrey, M. (2001). Model for the structure of the HIV gp41 ectodomain: insight into the intermolecular interactions of the gp41 loop. *Biochim. Biophys. Acta*. 1536(2-3):116–122.

[41] Zhou, G., Wu, D., Hermel, E., Balogh, E., and Gochin, M. (2010). Design, synthesis, and evaluation of indole compounds as novel inhibitors targeting Gp41. *Bioorg. Med. Chem. Lett*. 20(5):1500–1503.

[42] Yu, X., Lu, L., Cai, L., Tong, P., Tan, S., Zou, P., Meng, F., Chen, Y. h., and Jiang, S. (2011). Mutations of Gln64 in the HIV-1 gp41 N-Terminal Heptad Repeat Render Viruses Resistant to Peptide HIV Fusion Inhibitors Targeting the gp41 Pocket. *J. Virol*. 86(1):589–593.

[43] Ingallinella, P., Bianchi, E., Ladwa, N. A., Wang, Y.-J., Hrin, R., Veneziano, M., Bonelli, F., Ketas, T. J., Moore, J. P., Miller, M. D., and Pessi, A. (2009). Addition of a cholesterol group to an HIV-1 peptide fusion inhibitor dramatically increases its antiviral potency. *Proc. Natl. Acad. Sci. U. S. A.* 106(14):5801–5806.

[44] Singh, P., Sharma, P., Bisetty, K., Corcho, F. J., and Perez, J. J. (2011). Comparative structural studies of T-20 analogues using molecular dynamics. *Comp. Theor. Chem*. 974(1-3):122–132.

[45] Martins Do Canto, A. M. T., Palace Carvalho, A. J., Prates Ramalho, J. P., and Loura, L. M. S. (2008). T-20 and T-1249 HIV fusion inhibitors' structure and conformation in solution: a molecular dynamics study. *J. Pept. Sci*. 14(4):442–447.

[46] Martins do Canto, A. M. T., Carvalho, A. J. P., Ramalho, J. P. P., and Loura, L. M. S. (2010). Structure and conformation of HIV fusion inhibitor peptide T-1249 in presence of model membranes: A molecular dynamics study. *J. Mol. Struc. (Theochem.)*. 946(1-3):119–124.

[47] Martins do Canto, A. M. T., Carvalho, A. J. P., Ramalho, J. P. P., and Loura, L. M. S. (2011). Molecular dynamics simulations of T-20 HIV fusion inhibitor interacting with model membranes. *Biophys. Chem*. 159(2-3):275–286.

[48] Martins do Canto, A. M. T., Palace Carvalho, A. J., Prates Ramalho, J. P., and Loura, L. M. S. (2012). Molecular Dynamics Simulation of HIV Fusion Inhibitor T-1249: Insights on Peptide-Lipid Interaction. *Comput. Math. Methods. Med*. 2012:1–14.

[49] Martins do Canto, A., Carvalho, A., Ramalho, J., and Loura, L. (2013). Effect of Amphipathic HIV Fusion Inhibitor Peptides on POPC and POPC/Cholesterol Membrane Properties: A Molecular Simulation Study. *Int. J. Mol. Sci*. 14(7):14724–14743.

[50] Matos, P. M., Castanho, M. A. R. B., and Santos, N. C. (2010). HIV-1 Fusion Inhibitor Peptides Enfuvirtide and T-1249 Interact with Erythrocyte and Lymphocyte Membranes. *PLoS One*. 5(3):e9830.

[51] Hartono, Y. D., Lee, A. N., Lee-Huang, S., and Zhang, D. (2011). Computational study of bindings of HL9, a nonapeptide fragment of human lysozyme, to HIV-1 fusion protein gp41. *Bioorg. Med. Chem. Lett*. 21(6):1607–1611.

[52] Münch, J., Ständker, L., Adermann, K., Schulz, A., Schindler, M., Chinnadurai, R., Pöhlmann, S., Chaipan, C., Biet, T., Peters, T., Meyer, B., Wilhelm, D., Lu, H., Jing, W., Jiang, S., Forssmann, W.-G., and Kirchhoff, F. (2007). Discovery and Optimization of a Natural HIV-1 Entry Inhibitor Targeting the gp41 Fusion Peptide. *Cell*. 129(2):263–275.

[53] Gonzalez, E., Ballana, E., Clotet, B., and Esté, J. A. (2011). Development of resistance to VIR-353 with cross-resistance to the natural HIV-1 entry virus inhibitory peptide (VIRIP). *AIDS*. 25(13):1557–1583.

[54] González-Ortega, E., Ballana, E., Badia, R., Clotet, B., and Esté, J. A. (2011). Compensatory mutations rescue the virus replicative capacity of VIRIP-resistant HIV-1. *Antiviral Res*. 92(3):479–483.

[55] Gordon, L. M., Mobley, P. W., Pilpa, R., Sherman, M. A., and Waring, A. J. (2002). Conformational mapping of the N-terminal peptide of HIV-1 gp41 in membrane environments using (13)C-enhanced Fourier transform infrared spectroscopy. *Biochim. Biophys. Acta*. 1559(2):96–120.

[56] Gordon, L. M., Mobley, P. W., Lee, W., Eskandari, S., Kaznessis, Y. N., Sherman, M. A., and Waring, A. J. (2004). Conformational mapping of the N-terminal peptide of HIV-1 gp41 in lipid detergent and aqueous environments using 13C-enhanced Fourier transform infrared spectroscopy. *Protein Sci*. 13(4):1012–1030.

[57] Jaroniec, C. P., Kaufman, J. D., Stahl, S. J., Viard, M., Blumenthal, R., Wingfield, P. T., and Bax, A. (2005). Structure and Dynamics of Micelle-Associated Human Immunodeficiency Virus gp41 Fusion Domain. *Biochemistry*. 44(49):16167–16180.

[58] Li, Y. and Tamm, L. K. (2007). Structure and Plasticity of the Human Immunodeficiency Virus gp41 Fusion Domain in Lipid Micelles and Bilayers. *Biophys. J*. 93(3):876–885.

[59] Gomara, M. J., Lorizate, M., Huarte, N., Mingarro, I., Perez-Payá, E., and Nieva, J. L. (2006). Hexapeptides that interfere with HIV-1 fusion peptide activity in liposomes block GP41-mediated membrane fusion. *FEBS Letters*. 580(11):2561–2566.

[60] Herrera, E., Gomara, M. J., Mazzini, S., Ragg, E., and Haro, I. (2009). Synthetic Peptides of Hepatitis G Virus (GBV-C/HGV) in the Selection of Putative Peptide Inhibitors of the HIV-1 Fusion Peptide. *J. Phys. Chem. B*. 113(20):7383–7391.

[61] Herrera, E., Tenckhoff, S., Gomara, M. J., Galatola, R., Bleda, M. J., Gil, C., Ercilla, G., Gatell, J. M., Tillmann, H. L., and Haro, I. (2010). Effect of Synthetic Peptides Belonging to E2 Envelope Protein of GB Virus C on Human Immunodeficiency Virus Type 1 Infection. *J. Med. Chem*. 53(16):6054–6063.

[62] Koedel, Y., Eissmann, K., Wend, H., Fleckenstein, B., and Reil, H. (2011). Peptides Derived from a Distinct Region of GB Virus C Glycoprotein E2 Mediate Strain-Specific HIV-1 Entry Inhibition. *J. Virol*. 85(14):7037–7047.

[63] Sánchez-Martín, M. J., Hristova, K., Pujol, M., Gomara, M. J., Haro, I., Asunción Alsina, M., and Antònia Busquets, M. (2011). Analysis of HIV-1 fusion peptide inhibition by synthetic peptides from E1 protein of GB virus C. *J. Colloid Interface Sci*. 360(1):124–131.

[64] Sánchez-Martín, M. J., Busquets, M. A., Girona, V., Haro, I., Alsina, M. A., and Pujol, M. (2011). Effect of E1(64-81) hepatitis G peptide on the in vitro interaction of HIV-1 fusion peptide with membrane models. *Biochim. Biophys. Acta*. 1808(9):2178–2188.

[65] Haro, I., Gómara, M. J., Galatola, R., Domènech, O., Prat, J., Girona, V., and Busquets, M. A. (2011). Study of the inhibition capacity of an 18-mer peptide domain of GBV-C virus on gp41-FP HIV-1 activity. *Biochim. Biophys. Acta*. 1808(6):1567–1573.

[66] Fernández, L., Chan, W. C., Egido, M., Gomara, M. J., and Haro, I. (2012). Synthetic peptides derived from an N-terminal domain of the E2 protein of GB virus C in the study of GBV-C/HIV-1 co-infection. *J. Pept. Sci*. 18(5):326–335.

[67] Gallaher, W. R. (1987). Detection of a fusion peptide sequence in the transmembrane protein of human immunodeficiency virus. *Cell*. 50(3):327–328.

[68] Delahunty, M. D., Rhee, I., Freed, E. O., and Bonifacino, J. S. (1996). Mutational analysis of the fusion peptide of the human immunodeficiency virus type 1: identification of critical glycine residues. *Virology*. 218(1):94–102.

[69] Torres, O. and Bong, D. (2011). Determinants of Membrane Activity from Mutational Analysis of the HIV Fusion Peptide. *Biochemistry*. 50(23):5195–5207.

[70] Freed, E. O., Myers, D. J., and Risser, R. (1990). Characterization of the fusion domain of the human immunodeficiency virus type 1 envelope glycoprotein gp41. *Proc. Natl. Acad. Sci. U. S. A*. 87(12):4650–4654.

[71] Freed, E. O., Delwart, E. L., Buchschacher, G. L., and Panganiban, A. T. (1992). A mutation in the human immunodeficiency virus type 1 transmembrane glycoprotein gp41 dominantly interferes with fusion and infectivity. *Proc. Natl. Acad. Sci. U. S. A*. 89(1):70–74.

[72] Dimitrov, A. S., Rawat, S. S., Jiang, S., and Blumenthal, R. (2003). Role of the Fusion Peptide and Membrane-Proximal Domain in HIV-1 Envelope Glycoprotein-Mediated Membrane Fusion. *Biochemistry*. 42(48):14150–14158.

[73] Chang, D. K., Cheng, S. F., and Chien, W. J. (1997). The amino-terminal fusion domain peptide of human immunodeficiency virus type 1 gp41 inserts into the sodium dodecyl sulfate micelle primarily as a helix with a conserved glycine at the micelle-water interface. *J. Virol*. 71(9):6593–6602.

[74] Yang, J., Gabrys, C. M., and Weliky, D. P. (2001). Solid-State Nuclear Magnetic Resonance Evidence for an Extended $\beta$ Strand Conformation of the Membrane-Bound HIV-1 Fusion Peptide. *Biochemistry*. 40(27):8126–8137.

[75] Yang, J. and Weliky, D. P. (2003). Solid-State Nuclear Magnetic Resonance Evidence for Parallel and Antiparallel Strand Arrangements in the Membrane-Associated HIV-1 Fusion Peptide. *Biochemistry*. 42(40):11879–11890.

[76] Sackett, K. and Shai, Y. (2003). How structure correlates to function for membrane associated HIV-1 gp41 constructs corresponding to the N-terminal half of the ectodomain. *J. Mol. Biol*. 333(1):47–58.

[77] Yang, J., Prorok, M., Castellino, F. J., and Weliky, D. P. (2004). Oligomeric -Structure of the Membrane-Bound HIV-1 Fusion Peptide Formed from Soluble Monomers. *Biophys. J*. 87(3):1951–1963.

[78] Sackett, K. and Shai, Y. (2005). The HIV Fusion Peptide Adopts Intermolecular Parallel -Sheet Structure in Membranes when Stabilized by the Adjacent N-Terminal Heptad Repeat: A 13C FTIR Study. *J. Mol. Biol*. 350(4):790–805.

[79] Qiang, W., Yang, J., and Weliky, D. P. (2007). Solid-State Nuclear Magnetic Resonance Measurements of HIV Fusion Peptide to Lipid Distances Reveal the Intimate Contact of Strand Peptide with Membranes and the Proximity of the Ala-14Gly-16 Region with Lipid Headgroups. *Biochemistry*. 46(17):4997–5008.

[80] Qiang, W., Bodner, M. L., and Weliky, D. P. (2008). Solid-State NMR Spectroscopy of Human Immunodeficiency Virus Fusion Peptides Associated with Host-Cell-Like Membranes: 2D Correlation Spectra and Distance Measurements Support a Fully Extended Conformation and Models for Specific Antiparallel Strand Registries. *J. Am. Chem. Soc*. 130(16):5459–5471.

[81] Qiang, W., Sun, Y., and Weliky, D. P. (2009). A strong correlation between fusogenicity and membrane insertion depth of the HIV fusion peptide. *Proc. Natl. Acad. Sci. U. S. A*. 106(36):15314–15319.

[82] Schmick, S. D. and Weliky, D. P. (2010). Major Antiparallel and Minor Parallel Sheet Populations Detected in the Membrane-Associated Human Immunodeficiency Virus Fusion Peptide. *Biochemistry*. 49(50):10623–10635.

[83] Reichert, J., Grasnick, D., Afonin, S., Buerck, J., Wadhwani, P., and Ulrich, A. S. (2006). A critical evaluation of the conformational requirements of fusogenic peptides in membranes. *Eur. Biophys. J*. 36(4-5):405–413.

[84] Grasnick, D., Sternberg, U., Strandberg, E., Wadhwani, P., and Ulrich, A. S. (2011). Irregular structure of the HIV fusion peptide in membranes demonstrated by solid-state NMR and MD simulations. *Eur. Biophys. J*. 40(4):529–543.

[85] Martin, I., Schaal, H., Scheid, A., and Ruysschaert, J. M. (1996). Lipid membrane fusion induced by the human immunodeficiency virus type 1 gp41 N-terminal extremity is determined by its orientation in the lipid bilayer. *J. Virol*. 70(1):298–304.

[86] Peisajovich, S. G., Epand, R. F., Pritsker, M., Shai, Y., and Epand, R. M. (2000). The Polar Region Consecutive to the HIV Fusion Peptide Participates in Membrane Fusion. *Biochemistry*. 39(7):1826–1833.

[87] Saez-Cirion, A. and Nieva, J. (2002). Conformational transitions of membrane-bound HIV-1 fusion peptide. *Biochim. Biophys. Acta*. 1564(1):57–65.

[88] Castano, S. and Desbat, B. (2005). Structure and orientation study of fusion peptide FP23 of gp41 from HIV-1 alone or inserted into various lipid membrane models (mono-, bi- and multibi-layers) by FT-IR spectroscopies and Brewster angle microscopy. *Biochim. Biophys. Acta*. 1715(2):81–95.

[89] Morris, K. F., Gao, X., and Wong, T. C. (2004). The interactions of the HIV gp41 fusion peptides with zwitterionic membrane mimics determined by NMR spectroscopy. *Biochim. Biophys. Acta*. 1667(1):67–81.

[90] Barz, B., Wong, T. C., and Kosztin, I. (2008). Membrane curvature and surface area per lipid affect the conformation and oligomeric state of HIV-1 fusion peptide: a combined FTIR and MD simulation study. *Biochim. Biophys. Acta*. 1778(4):945–953.

[91] Buzón, V., Padrós, E., and Cladera, J. (2005). Interaction of Fusion Peptides from HIV gp41 with Membranes: A Time-Resolved Membrane Binding, Lipid Mixing, and Structural Study. *Biochemistry*. 44(40):13354–13364.

[92] Buzón, V. and Cladera, J. (2006). Effect of Cholesterol on the Interaction of the HIV GP41 Fusion Peptide with Model Membranes. Importance of the Membrane Dipole Potential. *Biochemistry*. 45(51):15768–15775.

[93] Zheng, Z., Yang, R., Bodner, M. L., and Weliky, D. P. (2006). Conformational Flexibility and Strand Arrangements of the Membrane-Associated HIV Fusion Peptide Trimer Probed by Solid-State NMR Spectroscopy. *Biochemistry*. 45(43):12960–12975.

[94] Sackett, K., Nethercott, M. J., Epand, R. F., Epand, R. M., Kindra, D. R., Shai, Y., and Weliky, D. P. (2010). Comparative Analysis of Membrane-Associated Fusion Peptide Secondary Structure and Lipid Mixing Function of HIV gp41 Constructs that Model the Early Pre-Hairpin Intermediate and Final Hairpin Conformations. *J. Mol. Biol*. 397(1):301–315.

[95] Tristram-Nagle, S., Chan, R., Kooijman, E., Uppamoochikkal, P., Qiang, W., Weliky, D. P., and Nagle, J. F. (2010). HIV Fusion Peptide Penetrates, Disorders, and Softens T-Cell Membrane Mimics. *J. Mol. Biol*. 402(1):139–153.

[96] Vogel, E. P., Curtis-Fisk, J., Young, K. M., and Weliky, D. P. (2011). Solid-State Nuclear Magnetic Resonance (NMR) Spectroscopy of Human Immunodeficiency Virus gp41 Protein That Includes the Fusion Peptide: NMR Detection of Recombinant Fgp41 in Inclusion Bodies in Whole Bacterial Cells and Structural Characterization of Purified and Membrane-Associated Fgp41. *Biochemistry*. 50(46):10013–10026.

[97] Lai, A. L., Moorthy, A. E., Li, Y., and Tamm, L. K. (2012). Fusion Activity of HIV gp41 Fusion Domain Is Related to Its Secondary Structure and Depth of Membrane Insertion in a Cholesterol-Dependent Fashion. *J. Mol. Biol*. 418(1-2):3–15.

[98] Gabrys, C. M., Qiang, W., Sun, Y., Xie, L., Schmick, S. D., and Weliky, D. P. (2013). Solid-State Nuclear Magnetic Resonance Measurements of HIV Fusion Peptide (13)CO to Lipid (31)P Proximities Support Similar Partially Inserted Membrane Locations of the $\alpha$ Helical and $\beta$ Sheet Peptide Structures. *J. Phys. Chem. A*. 117(39):9848–9859.

[99] Gordon, L. M., Nisthal, A., Lee, A. B., Eskandari, S., Ruchala, P., Jung, C.-L., Waring, A. J., and Mobley, P. W. (2008). Structural and functional properties of peptides based on the N-terminus of HIV-1 gp41 and the C-terminus of the amyloid-beta protein. *Biochim. Biophys. Acta*. 1778(10):2127–2137.

[100] Zhan, H. and Lazaridis, T. (2012). Influence of the membrane dipole potential on peptide binding to lipid bilayers. *Biophys. Chem*. 161:1–7.

[101] Bradshaw, J. P., Darkes, M. J. M., Harroun, T. A., Katsaras, J., and Epand, R. M. (2000). Oblique Membrane Insertion of Viral Fusion Peptide Probed by Neutron Diffraction. *Biochemistry*. 39(22):6581–6585.

[102] Charloteaux, B., Lorin, A., Brasseur, R., and Lins, L. (2009). The "Tilted Peptide Theory" Links Membrane Insertion Properties and Fusogenicity of Viral Fusion Peptides. *Protein Pept. Lett*. 16(7):718–725.

[103] Charloteaux, B., Lorin, A., Crowet, J. M., Stroobant, V., Lins, L., Thomas, A., and Brasseur, R. (2006). The N-terminal 12 Residue Long Peptide of HIV gp41 is the Minimal Peptide Sufficient to Induce Significant T-cell-like Membrane Destabilization in Vitro. *J. Mol. Biol*. 359(3):597–609.

[104] Cheng, Y., Li, D., Ji, B., Shi, X., and Gao, H. (2010). Structure-based design of carbon nanotubes as HIV-1 protease inhibitors: atomistic and coarse-grained simulations. *J. Mol. Graph. Model*. 29(2):171–177.

[105] Lins, L., Decaffmeyer, M., Thomas, A., and Brasseur, R. (2008). Relationships between the orientation and the structural properties of peptides and their membrane interactions. *Biochim. Biophys. Acta*. 1778(7-8):1537–1544.

[106] Taylor, A. and Sansom, M. S. P. (2010). Studies on viral fusion peptides: the distribution of lipophilic and electrostatic potential over the peptide determines the angle of insertion into a membrane. *Eur. Biophys. J*. 39(11):1537–1545.

[107] Shchelokovskyy, P., Tristram-Nagle, S., and Dimova, R. (2011). Effect of the HIV-1 fusion peptide on the mechanical properties and leaflet coupling of lipid bilayers. *New J. Phys*. 13(2):025004.

[108] Bitler, A., Lev, N., Fridmann-Sirkis, Y., Blank, L., Cohen, S. R., and Shai, Y. (2010). Kinetics of interaction of HIV fusion protein (gp41) with lipid membranes studied by real-time AFM imaging. *Ultramicroscopy*. 110(6):694–700.

[109] Ivankin, A., Kuzmenko, I., and Gidalevitz, D. (2012). Cholesterol Mediates Membrane Curvature during Fusion Events. *Phys. Rev. Lett*. 108(23):238103.

[110] Venken, T., Voet, A., De Maeyer, M., De Fabritiis, G., and Sadiq, S. K. (2013). Rapid Conformational Fluctuations of Disordered HIV-1 Fusion Peptide in Solution. *J. Chem. Theory Comput*. 9(7):2870–2874.

[111] Wong, T. C. (2003). Membrane structure of the human immunodeficiency virus gp41 fusion peptide by molecular dynamics simulation. II. The glycine mutants. *Biochim. Biophys. Acta*. 1609(1):45–54.

[112] Kliger, Y., Aharoni, A., Rapaport, D., Jones, P., Blumenthal, R., and Shai, Y. (1997). Fusion peptides derived from the HIV type 1 glycoprotein 41 associate within phospholipid membranes and inhibit cell-cell fusion - Structure-function study. *J. Biol. Chem*. 272(21):13496–13505.

[113] Promsri, S., Ullmann, G. M., and Hannongbua, S. (2012). Molecular dynamics simulation of HIV-1 fusion domain-membrane complexes: Insight into the N-terminal gp41 fusion mechanism. *Biophys. Chem*. 170:9–16.

[114] Pollard, V. W. and Malim, M. H. (1998). The HIV-1 rev protein. *Annu. Rev. Microbiol*. 52:491–532.

[115] Suhasini, M. and Reddy, T. R. (2009). Cellular proteins and HIV-1 Rev function. *Curr. HIV Res*. 7(1):91–100.

[116] Fukuda, M., Asano, S., Nakamura, T., Adachi, M., Yoshida, M., Yanagida, M., and Nishida, E. (1997). CRM1 is responsible for intracellular transport mediated by the nuclear export signal. *Nature*. 390(6657):308–311.

[117] Kubota, S. and Pomerantz, R. J. (1998). A cis-acting peptide signal in human immunodeficiency virus type I Rev which inhibits nuclear entry of small proteins. *Oncogene*. 16(14):1851–1861.

[118] Henderson, B. R. and Percipalle, P. (1997). Interactions between HIV Rev and nuclear import and export factors: the Rev nuclear localisation signal mediates specific binding to human importin-$\beta$. *J. Mol. Biol*. 274(5):693–707.

[119] Fernandes, J., Jayaraman, B., and Frankel, A. (2012). The HIV-1 rev response element: An RNA scaffold that directs the cooperative assembly of a homo-oligomeric ribonucleoprotein complex. *RNA Biol*. 9(1):4–9.

[120] Daly, T. J., Doten, R. C., Rusche, J. R., and Auer, M. (1995). The amino terminal domain of HIV-1 Rev is required for discrimination of the RRE from nonspecific RNA. *J. Mol. Biol*. 253(2):243–258.

[121] Sloan, E. A., Kearney, M. F., Gray, L. R., Anastos, K., Daar, E. S., Margolick, J., Maldarelli, F., Hammarskjold, M. L., and Rekosh, D. (2013). Limited Nucleotide Changes in the Rev Response Element (RRE) during HIV-1 Infection Alter Overall Rev-RRE Activity and Rev Multimerization. *J. Virol*. 87(20):11173–11186.

[122] Kumar, S., Bose, D., Suryawanshi, H., Sabharwal, H., Mapa, K., and Maiti, S. (2011). Specificity of RSG-1.2 Peptide Binding to RRE-IIB RNA Element of HIV-1 over Rev Peptide Is Mainly Enthalpic in Origin. *PLoS One*. 6(8):e23300.

[123] Wilkinson, T. A., Botuyan, M. V., Kaplan, B. E., Rossi, J. J., and Chen, Y. (2000). Arginine side-chain dynamics in the HIV-1 Rev-RRE complex. *J. Mol. Biol*. 303(4):515–529.

[124] Michael, L. A., Chenault, J. A., Miller, B. R., III, Knolhoff, A. M., and Nagan, M. C. (2009). Water, Shape Recognition, Salt Bridges, and Cation–Pi Interactions Differentiate Peptide Recognition of the HIV Rev-Responsive Element. *J. Mol. Biol*. 392(3):774–786.

[125] Casu, F., Duggan, B. M., and Hennig, M. (2013). The Arginine-Rich RNA-Binding Motif of HIV-1 Rev Is Intrinsically Disordered and Folds upon RRE Binding. *Biophys. J*. 105(4):1004–1017.

[126] Fischer, U., Huber, J., Boelens, W. C., Mattaj, I. W., and Lührmann, R. (1995). The HIV-1 Rev activation domain is a nuclear export signal that accesses an export pathway used by specific cellular RNAs. *Cell*. 82(3):475–483.

[127] Hakata, Y., Yamada, M., Mabuchi, N., and Shida, H. (2002). The Carboxy-Terminal Region of the Human Immunodeficiency Virus Type 1 Protein Rev Has Multiple Roles in Mediating CRM1-Related Rev Functions. *J. Virol*. 76(16):8079–8089.

[128] Güttler, T., Madl, T., Neumann, P., Deichsel, D., Corsini, L., Monecke, T., Ficner, R., Sattler, M., and Görlich, D. (2010). NES consensus redefined by structures of PKI-type and Rev-type nuclear export signals bound to CRM1. *Nat. Struct. Mol. Biol*. 17(11):1367–1376.

[129] Fang, J., Kubota, S., Yang, B., Zhou, N., Zhang, H., Godbout, R., and Pomerantz, R. J. (2004). A DEAD box protein facilitates HIV-1 replication as a cellular co-factor of Rev. *Virology*. 330(2):471–480.

[130] Edgcomb, S. P., Carmel, A. B., Naji, S., Ambrus-Aikelin, G., Reyes, J. R., Saphire, A. C. S., Gerace, L., and Williamson, J. R. (2012). DDX1 Is an RNA-Dependent ATPase Involved in HIV-1 Rev Function and Virus Replication. *J. Mol. Biol*. 415(1):61–74.

[131] Robertson-Anderson, R. M., Wang, J., Edgcomb, S. P., Carmel, A. B., Williamson, J. R., and Millar, D. P. (2011). Single-Molecule Studies Reveal that DEAD Box Protein DDX1 Promotes Oligomerization of HIV-1 Rev on the Rev Response Element. *J. Mol. Biol*. 410(5):959–971.

[132] Yedavalli, V. S. R. K., Neuveut, C., Chi, Y.-H., Kleiman, L., and Jeang, K.-T. (2004). Requirement of DDX3 DEAD box RNA helicase for HIV-1 Rev-RRE export function. *Cell*. 119(3):381–392.

[133] Zhou, X., Luo, J., Mills, L., Wu, S., Pan, T., Geng, G., Zhang, J., Luo, H., Liu, C., and Zhang, H. (2013). DDX5 facilitates HIV-1 replication as a cellular co-factor of Rev. *PLoS One*. 8(5):e65040.

[134] Yasuda-Inoue, M., Kuroki, M., and Ariumi, Y. (2013). Distinct DDX DEAD-box RNA helicases cooperate to modulate the HIV-1 Rev function. *Biochem. Biophys. Res. Commun*. 434(4):803–808.

[135] Daly, T. J., Cook, K. S., Gray, G. S., Maione, T. E., and Rusche, J. R. (1989). Specific binding of HIV-1 recombinant Rev protein to the Rev-responsive element in vitro. *Nature*. 342(6251):816–819.

[136] Daly, T. J., Doten, R. C., Rennert, P., Auer, M., Jaksche, H., Donner, A., Fisk, G., and Rusche, J. R. (1993). Biochemical characterization of binding of multiple HIV-1 Rev monomeric proteins to the Rev responsive element. *Biochemistry*. 32(39):10497–10505.

[137] Mann, D. A., Mikaélian, I., Zemmel, R. W., Green, S. M., Lowe, A. D., Kimura, T., Singh, M., Butler, P. J., Gait, M. J., and Karn, J. (1994). A molecular rheostat. Co-operative rev binding to stem I of the rev-response element modulates human immunodeficiency virus type-1 late gene expression. *J. Mol. Biol*. 241(2):193–207.

[138] Daugherty, M. D., Booth, D. S., Jayaraman, B., Cheng, Y., and Frankel, A. D. (2010). HIV Rev response element (RRE) directs assembly of the Rev homooligomer into discrete asymmetric complexes. *Proc. Natl. Acad. Sci. U. S. A*. 107(28):12481–12486.

[139] DiMattia, M. A., Watts, N. R., Stahl, S. J., Rader, C., Wingfield, P. T., Stuart, D. I., Steven, A. C., and Grimes, J. M. (2010). Implications of the HIV-1 Rev dimer structure at 3.2 Å resolution for multimeric binding to the Rev response element. *Proc. Natl. Acad. Sci. U. S. A*. 107(13):5810–5814.

[140] Daugherty, M. D., Liu, B., and Frankel, A. D. (2010). Structural basis for cooperative RNA binding and export complex assembly by HIV Rev. *Nat. Struct. Mol. Biol*. 17(11):1337–1342.

[141] Jain, C. and Belasco, J. G. (2001). Structural model for the cooperative assembly of HIV-1 Rev multimers on the RRE as deduced from analysis of assembly-defective mutants. *Mol. Cell*. 7(3):603–614.

[142] Daugherty, M. D., D'Orso, I., and Frankel, A. D. (2008). A Solution to Limited Genomic Capacity: Using Adaptable Binding Surfaces to Assemble the Functional HIV Rev Oligomer on RNA. *Mol. Cell*. 31(6):824–834.

[143] Edgcomb, S. P., Aschrafi, A., Kompfner, E., Williamson, J. R., Gerace, L., and Hennig, M. (2008). Protein structure and oligomerization are important for the formation of export-competent HIV-1 Rev-RRE complexes. *Protein Sci*. 17(3):420–430.

[144] Auer, M., Gremlich, H. U., Seifert, J. M., Daly, T. J., Parslow, T. G., Casari, G., and Gstach, H. (1994). Helix-loop-helix motif in HIV-1 Rev. *Biochemistry*. 33(10):2988–2996.

[145] Pond, S. J. K., Ridgeway, W. K., Robertson, R., Wang, J., and Millar, D. P. (2009). HIV-1 Rev protein assembles on viral RNA one molecule at a time. *Proc. Natl. Acad. Sci. U. S. A*. 106(5):1404–1408.

[146] Daelemans, D., Costes, S. V., Cho, E. H., Erwin-Cohen, R. A., Lockett, S., and Pavlakis, G. N. (2004). In vivo HIV-1 Rev multimerization in the nucleolus and cytoplasm identified by fluorescence resonance energy transfer. *J. Biol. Chem*. 279(48):50167–50175.

[147] Pallesen, J., Dong, M., Besenbacher, F., and Kjems, J. (2009). Structure of the HIV-1 Rev response element alone and in complex with regulator of virion (Rev) studied by atomic force microscopy. *FEBS Journal*. 276(15):4223–4232.

[148] Živković, J., Janssen, L., Alvarado, F., Speller, S., and Heus, H. A. (2012). Force spectroscopy of Rev-peptide–RRE interaction from HIV-1. *Soft Matter*. 8(7):2103–2109.

[149] Havlin, R. H., Blanco, F. J., and Tycko, R. (2007). Constraints on Protein Structure in HIV-1 Rev and RevRNA Supramolecular Assemblies from Two-Dimensional Solid State Nuclear Magnetic Resonance. *Biochemistry*. 46(11):3586–3593.

[150] Hammarskjold, M.-L. and Rekosh, D. (2011). A Long-Awaited Structure Is Rev-ealed. *Viruses*. 3(12):484–492.

[151] Malim, M. H. and Cullen, B. R. (1991). HIV-1 structural gene expression requires the binding of multiple Rev monomers to the viral RRE: implications for HIV-1 latency. *Cell*. 65(2):241–248.

[152] Malim, M. H., Böhnlein, S., Hauber, J., and Cullen, B. R. (1989). Functional dissection of the HIV-1 Rev trans-activator–derivation of a trans-dominant repressor of Rev function. *Cell*. 58(1):205–214.

[153] Chapman, R. L., Stanley, T. B., Hazen, R., and Garvey, E. P. (2002). Small molecule modulators of HIV Rev/Rev response element interaction identified by random screening. *Antiviral Res*. 54(3):149–162.

[154] Mills, N. L., Daugherty, M. D., Frankel, A. D., and Guy, R. K. (2006). An $\alpha$-Helical Peptidomimetic Inhibitor of the HIV-1 RevRRE Interaction. *J. Am. Chem. Soc*. 128(11):3496–3497.

[155] Shuck-Lee, D., Chen, F. F., Willard, R., Raman, S., Ptak, R., Hammarskjold, M. L., and Rekosh, D. (2008). Heterocyclic Compounds That Inhibit Rev-RRE Function and Human Immunodeficiency Virus Type 1 Replication. *Antimicrob. Agents Chemother*. 52(9):3169–3179.

[156] Wolff, B., Sanglier, J.-J., and Wang, Y. (1997). Leptomycin B is an inhibitor of nuclear export: inhibition of nucleo-cytoplasmic translocation of the human immunodeficiency virus type 1 (HIV-1) Rev protein and Rev-dependent mRNA. *Chem. Biol*. 4(2):139–147.

[157] Daelemans, D., Afonina, E., Nilsson, J., Werner, G., Kjems, J., De Clercq, E., Pavlakis, G. N., and Vandamme, A.-M. (2002). A synthetic HIV-1 Rev inhibitor interfering with the CRM1-mediated nuclear export. *Proc. Natl. Acad. Sci. U. S. A*. 99(22):14440–14445.

[158] Van Neck, T., Pannecouque, C., Vanstreels, E., Stevens, M., Dehaen, W., and Daelemans, D. (2008). Inhibition of the CRM1-mediated nucleocytoplasmic transport by N-azolylacrylates: Structure–activity relationship and mechanism of action. *Bioorg. Med. Chem*. 16(21):9487–9497.

[159] Shuck-Lee, D., Chang, H., Sloan, E. A., Hammarskjold, M. L., and Rekosh, D. (2011). Single-Nucleotide Changes in the HIV Rev-Response Element Mediate Resistance to Compounds That Inhibit Rev Function. *J. Virol*. 85(8):3940–3949.

[160] Vercruysse, T., Pardon, E., Vanstreels, E., Steyaert, J., and Daelemans, D. (2010). An Intrabody Based on a Llama Single-domain Antibody Targeting the N-terminal $\alpha$-Helical Multimerization Domain of HIV-1 Rev Prevents Viral Production. *J. Biol. Chem*. 285(28):21768–21780.

[161] Vercruysse, T., Pawar, S., De Borggraeve, W., Pardon, E., Pavlakis, G. N., Pannecouque, C., Steyaert, J., Balzarini, J., and Daelemans, D. (2011). Measuring cooperative Rev protein-protein interactions on Rev responsive RNA by fluorescence resonance energy transfer. *RNA Biol*. 8(2):316–324.

[162] Vercruysse, T., Boons, E., Venken, T., Vanstreels, E., Voet, A., Steyaert, J., De Maeyer, M., and Daelemans, D. (2013). Mapping the Binding Interface between an HIV-1 Inhibiting Intrabody and the Viral Protein Rev. *PLoS One*. 8(4):e60259.

# Chapter 2

## Molecular Dynamics simulations

"I feel so close to you right now, it's a force field"

Feel so Close - Calvin Harris

## 2.1 Introduction

Biomolecules can be seen as the machinery required by nature to survive, communicate and reproduce. Proteins are a very diverse class of macromolecules that perform numerous roles in living cells, such as regulation (e.g. in signalling pathways to transmit responses through the body), structure (e.g. collagen, which offers rigidity to fibrous tissues), transport of small molecules (e.g. water molecules by aquaporins) or catalyse reactions (e.g. the breakdown of metabolites), to name a few. Other biomolecules such as carbohydrates are essential as an energy source for cells (e.g. glycogen in muscle cells) or they play a role in the immune system (e.g. recognition of pathogens), while lipids are essential to give shape and compartmentalise different organelles from each other. Finally, nucleotides contain the genetic information essential for replication of the cell.

In biophysical research, a wide variety of techniques is available to study the role of these biomolecules. For example, X-ray crystallography can provide a highly detailed picture of the atomic positions of a biomolecular structure. However, a picture is often not sufficient. In those cases, a better option would be to construct a dynamic movie of an entire process. To use a real-life example, just as a picture of a running horse tells little about its pace (think of the movies of running horses made by Eadweard Muybridge in the 19th century [1]), a single protein conformation tells little about its dynamics (see Figure 2.1 for an illustration). To use another example: if we want to grasp how a ligand binds to a protein, we are not only interested in the final bound state but also in the entire binding event. Atoms and molecules collide with each other very frequently in the crowded

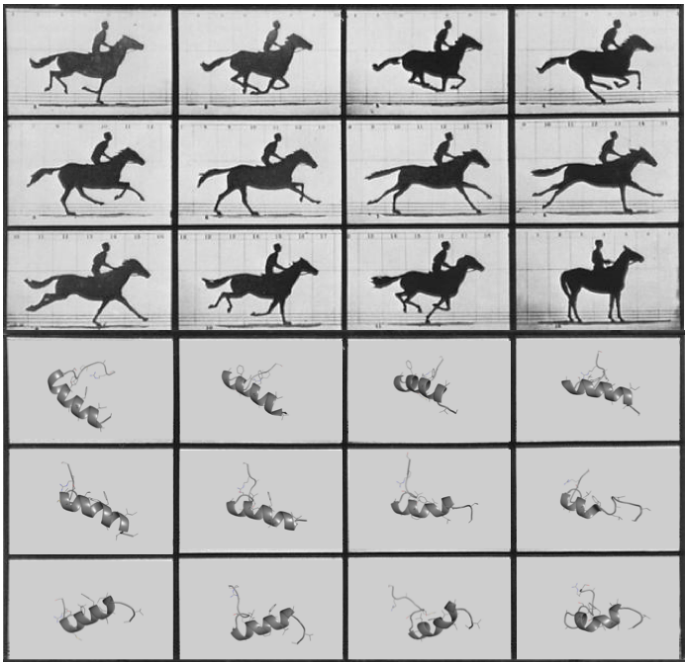

Figure 2.1: **An illustration of the relevance of dynamics.** Top: Sallie Gardner at a Gallop [1]. Eadward Muybridge conducted this photographic experiment in 1872 by taking several photographs of a galloping horse to find out if all feet where completely off the ground while trotting. This was, at that time, an unsolved mystery, as the human eye cannot distinguish the rapid movements of the horse. By converting the individual pictures into a movie, the horse was indeed captured with all feet off the ground during gallop. Bottom: gp41 FP in solution. In comparison, scientific arguments are often not explained by a single figure as well, as sometimes a series of pictures is necessary to fully understand and explain biomolecular questions. The depicted snapshots are gp41 FP conformations in solution based on PDB code 1ERF and where taken in 1.2 ns in 100 ps intervals, showing partial breakdown of the $\alpha$-helical conformation in a relatively short time.

environment of the living cell. Proteins, for example, are inherently flexible and the protein dynamics strongly influence the function of the protein itself. A detailed understanding of these interactions can explain the behaviour of larger macromolecules, or as R. Feynman has stated in a now famous quote: *"Everything that living things do can be understood in terms of the jigglings and wigglings of atoms"* [2, 3]. A detailed movie of molecular interactions will help us to understand the biological processes in living cells, also in those cases when mistakes lead to diseases. For example, when two viral proteins interact with each other and this interaction is essential for the replication of the virus, molecular modelling can aid to (a) investigate the dynamic nature of the interaction, and (b) find a way to inhibit this protein-protein interaction, thereby inhibiting the replication of the virus as well.

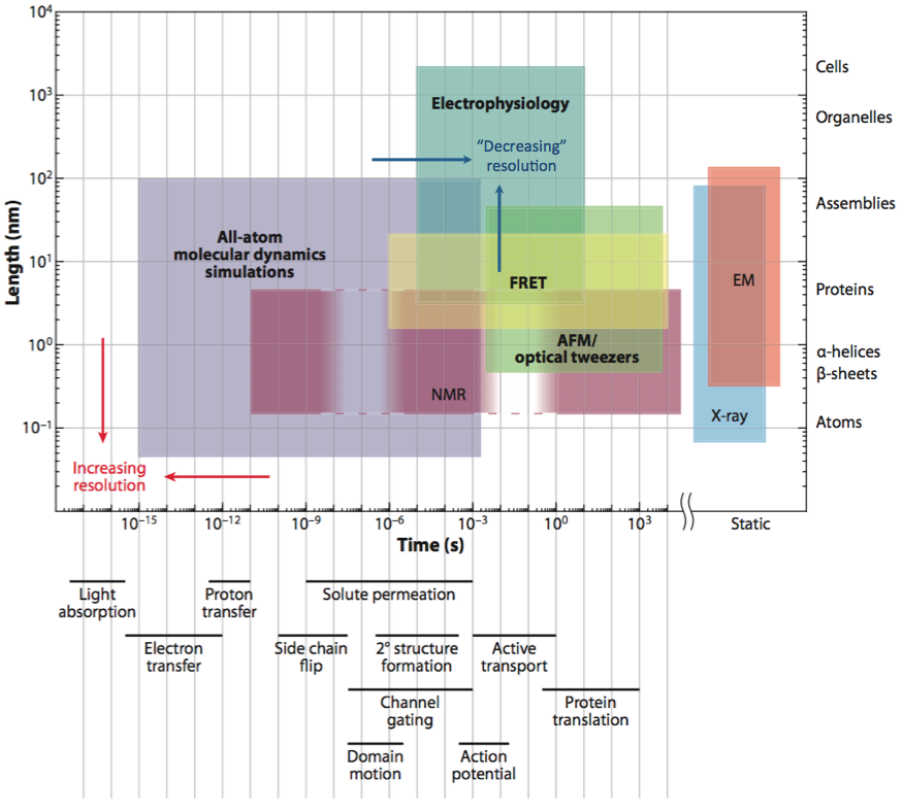


Figure 2.2: **Comparison of techniques to study biological processes and entities** in different time scales (x-axis) and of different sizes (y-axis). To optimise experimental methods, often an increase in resolution is required. In contrast, MD simulations struggle to "decrease" the resolution, that is, to study larger biomolecular structures at longer time scales. Hence, a combination of experimental and modelling techniques is usually required to solve biomolecular questions. Figure adapted from Dror *et al.* [4]

To allow a dynamical microscopic view of the biomolecules from living cells, techniques can be employed like NMR, electron cryo-microscopy and fluorescence based methods such as FRET. While those methods provide a valuable understanding of macromolecular structures, they are often limited by both spatial and temporal resolution [4]. In other words, biophysical techniques are not always capable of studying small systems or fast processes at a highly detailed atomic level. Attempts are being made to overcome the limitations, though these usually require more expensive scientific machinery. As such, molecular dynamics (MD) simulations can aid to trace the motions of biomolecules at an atomic scale (see Figure 2.2 for a comparison with experimental techniques).


Figure 2.3: **Demonstration of ensemble deception.** Experimental techniques often study ensembles in bulk, but the measured results can hide potentially interesting distinct microstates. Just as taking a picture with a long exposure time can blur the motions of the individual cars on a highway, so can the motions of individual proteins be hidden in ensemble measurements, which can potentially lead to faulty conclusions. Note that while averaging over distinct populations can be deceiving, conversely MD simulations can be inaccurate as well when only a limited amount of conformational space is sampled. Picture reproduced with permission of © Steven Duerinckx www.darkink.be [5].

In a MD simulation, the positions and velocities of atoms are calculated using classical Newtonian physics to predict the motions of biomolecular systems. To define the forces of all the atoms in a system, an empirical force field is used based on parameter fitting of quantum chemical and/or experimental data. While electrons are explicitly modelled in quantum mechanics, atoms are represented as rigid spheres connected by unbreakable bonds in a force field.

Nowadays, when starting from a reliable experimental structure or model and considering the limitations of the chosen force field, MD simulations are considered as a "virtual microscope". In addition, the behaviour of a single protein can be tracked, which is not always feasible with experimental techniques where often an ensemble of molecules is studied in bulk. Those ensemble methods can mask infrequent but possibly important molecular processes or interactions of a protein. As such, utilising single molecule techniques like MD allows exposure of dynamical events that are often averaged in ensemble measurements (see Figure 2.3 for an illustration of this principle).

MD simulation is a highly multidisciplinary technique, combining aspects from physics, mathematics and informatics to solve biology and chemistry questions

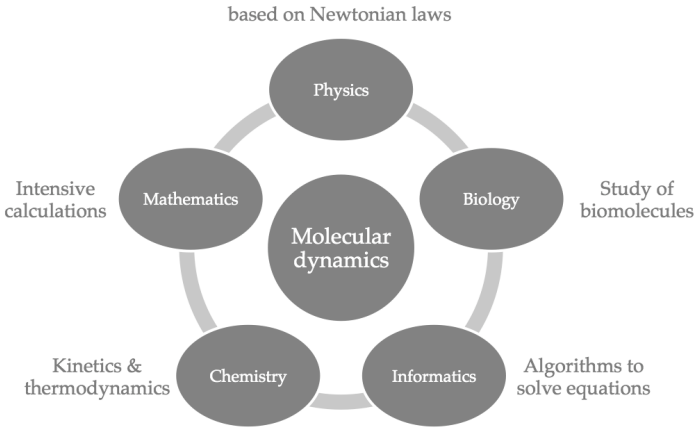


Figure 2.4: **Disciplines in MD simulations.** In MD simulations, many different fields are combined to solve biomolecular questions.

(see Figure 2.4). Numerous MD packages like GROMACS (GROningen Machine for Chemical Simulations) [6], AMBER (Assisted Model Building with Energy Refinement) [7], CHARMM (Chemistry at HARvard Molecular Mechanics) [8], and NAMD (Not just Another Molecular Dynamics program) [9] are becoming increasingly popular [10]. In parallel, a number of protein optimised force fields such as the AMBER and CHARMM force fields (not to be confused with the package name) have been developed.

A wide number of review articles have been published outlining the general principles of MD simulations [11–16], its applications in drug design [17–22], the folding of proteins [17, 23–26] and current challenges [27–29], like assessing longer time scales [4, 30–32]. In addition, researchers are becoming aware that protein flexibility needs to be taken into account for structure based drug design [33, 34], for example in virtual screening protocols [35–39].

Below, we will first outline the general principles of MD simulations and its implementations. Next, we list a number of challenges and limitations of the method.

## 2.2 Theoretical principles of Molecular Mechanics

In theory, the well-known Schrödinger's equation can be solved using quantum mechanical calculations to predict the properties of a given system *ab initio*. Unfortunately, applying those calculations on large proteins is impossible to track

the time scales scientists are interested in. This is even impossible for very small systems, so a number of approximations are implemented in standard MD simulations (and molecular mechanics in general).

First, the Born-Oppenheimer approximation assumes that nuclei are infinitely heavier than electrons. By decoupling motions of electrons and nuclei, the calculations are simplified significantly. In MD, the electronic motions are not considered explicitly and they depend on the position of the atoms in the system, so each atom is considered as a point mass.

Second, it is assumed that nucleic motions can behave classically according to Newton's second law, so quantum effects are not taken into account explicitly.

To obtain the sum of all forces, a third approximation is used, namely the application of an empirical force field, which we will explain more in detail below.

### 2.2.1 Force fields

In molecular mechanics, Newtonian laws are used instead of the Schrödinger's equation to describe a system. This yields a sufficiently reliable approximation of the total energy in a system at a much lower computational cost. The force field is the core of a MD simulation and determines the behaviour of the system of interest. Consequently, a vast number of force field have been developed specifically for the simulation of proteins. However, the systems' potential energy is not sufficient as such, but a combination of potential energy functions with an appropriate parameter set is required. In short, the force field is basically the collection of empirical potential functions that describe the interactions between all the atoms in the system. In other words, the sum of all the individual energy contributions in and between atoms constitutes the total potential energy $U$. Potential functions are bonded (i.e. internal interactions between atoms through covalent bonds, $E_b$) or non-bonded (i.e. external interactions by atoms that are not bonded, $E_{nb}$):

$$U = E_b + E_{nb} \tag{2.1}$$

$$E_b = E_{bond} + E_{angle} + E_{dihedral} \tag{2.2}$$

$$E_{nb} = E_{vdw} + E_{coul} \tag{2.3}$$

In AMBER [40], a force field frequently used in this thesis, the bonded and non-bonded terms consist of:

$$E_{bond} = \sum_{bonds} k_r(r - r_{eq}) \tag{2.4}$$

$$E_{angle} = \sum_{angles} k_\theta(\theta - \theta_{eq}) \tag{2.5}$$

$$E_{dihedral} = \sum_{dihedrals} k_\phi(1 + cos(n\phi - \gamma)) \tag{2.6}$$

$$E_{vdw} = \sum_{i<j} (\frac{A_{ij}}{{R_{ij}}^{12}} - \frac{B_{ij}}{{R_{ij}}^{6}}) \tag{2.7}$$

$$E_{coul} = \sum_{i<j} \frac{q_i q_j}{\epsilon R_{ij}} \tag{2.8}$$

A ball-and-spring-like model represents the molecules in a force field, where atoms are treated as charged point masses connected by bonds. In the force field described above, the energy terms between covalently bonded atoms are defined by bond stretching $E_{bond}$, an angle bending strain $E_{angle}$ and a dihedral term $E_{dihedral}$ for four steric atoms. Non-covalently bonded terms consist of a van der Waals term $E_{vdw}$, which defines atom-atom repulsion and dispersion interactions, and an electrostatic term $E_{coul}$, defined by Coulomb's law.

### Bonded terms

Chemical bonds and atomic angles are treated by simple harmonic springs, while dihedral or torsional rotations are defined by a sinusoidal term that corresponds to energy differences between eclipsed and staggered conformations. The bond term $E_{bond}$ is used to account for deformations in the bond length. Here, $k_r$ is the force constant, $r$ is the current bond length and $r_{eq}$ the bond length at equilibrium between two atoms. Similarly, the angle term consists of an angle force constant $k_\theta$, a current angle $\theta$ and an ideal angle $\theta_{eq}$ formed by three atoms. The four-atom dihedral potential contains an improper force constant $k_\phi$ with periodicity $n$, a phase shift $\gamma$ and the current torsional angle $\phi$. All the parameters result in a "fixed" list of atoms, that is, bonds are predefined and cannot be broken during the simulation.

### Non-bonded terms

For the non-bonded terms, repulsion and dispersion interactions are described by the van der Waals term $E_{vdw}$ with the Lennard-Jones (LJ) 6-12 potential as a function of the distance between two interacting atoms $r_{ij}$. A repulsive force exists at short atom-atom distances (defined by $A_{ij}$), while attractive London dispersion forces occur at longer distances (defined by $B_{ij}$). The constants $A_{ij}$ and $B_{ij}$ are determined experimentally and are specific for each atom pair. The electrostatic term $E_{coul}$ is treated by Coulomb's law and describes the interaction force between two static point charges. Here, $q_i$ and $q_j$ are the (usually quantum chemically derived) partial charges and $\epsilon$ is the permittivity of the medium.

The number of pairwise non-bonded interactions scales quadratically with the number of atoms in the system. Therefore, the amount of non-bonded pairs in the simulation is computed from a neighbour list. That is a list of all the non-bonded atoms within a certain radius. The list is updated regularly after a specified amount of time steps in the MD simulation (typically 10 ps). For the LJ interaction, the radius is often chosen between 1.0 and 1.4 nm. The long-range electrostatic interactions are notoriously the computationally most expensive part of a MD simulation step. Unfortunately, an abrupt cut-off generates artefacts in the simulation system as the Coulombic potential decays slowly at long distances. However, we cannot simply ignore these processes, as long-range electrostatics are essential for the structural stability of proteins, the folding of proteins [41] and in ligand binding events. Therefore, the particle mesh Ewald (PME) summation scheme has been developed to handle long-range electrostatic interactions more efficiently [42, 43]. The Ewald method partitions the potential into a short- and long-range part and sums these separately. The short part is calculated in real space while a Fourier transformed space is used to treat the long-range interactions. A more detailed implementation is found in the GROMACS manual [44].

### Parameterisation

Force field parameters are derived either from fitting to experimental data or from quantum mechanical calculations such as *ab initio* or semi-empirical methods. There is no unique way to implement a force field as exemplified by a wide range of different force fields developed by a variety of groups. The most common force fields for biomolecular simulations are AMBER, CHARMM, OPLS (Optimised Potentials for Liquid Simulations) and GROMOS (GROningen MOlecular Simulation). In a force field, the specific force parameters and ideal values must be parameterised with great care, as even small deviations may influence the entire protein dynamics. As an illustration, the small hydrophobin protein contains 70

amino acids, consisting of 982 atoms interacting with each other [45]. Due to the large amount of possible interactions, solving these equations is a computationally intensive task. Even in such a relatively small protein, there are 993 bonds, 1821 angles, 2835 dihedrals and 2629 pairwise interactions to be calculated in every single time step. Thus, a static picture of a protein can be deceiving as the molecule will in reality behave more like a giant ball of oscillating springs. It is with MD simulations that we try to solve this giant scientific puzzle in a step-by-step approach.

### 2.2.2 Integration

In a MD simulation, the time evolution of a system comprising N particles (with i = 1 ... N) is written as:

$$F_i = m_i a_i \tag{2.9}$$

Here, the force $F_i$ acting on a particle $i$ is defined by its mass $m_i$ and its acceleration $a_i$, which is the second derivative of the atomic position $r_i$ with time $t$, while velocity $v_i$ is the first derivative:

$$F_i = m_i \frac{\delta v_i}{\delta t} = m_i \frac{\delta^2 r_i}{\delta t^2} \tag{2.10}$$

The potential energy in the system $U$ is a function of the coordinates of the particles $x_i$ in the system:

$$F_i = -\frac{\delta U}{\delta r_i} \tag{2.11}$$

The force calculated here is simply the negative derivative of the potential energy. Essentially, the potential energy $U$ defines how the system evolves in time. For this reason, we need a set of initial coordinates ($x_i$) and assigned velocities ($v_i$) to start a MD simulation at time $i$. Next, the Newton second law is applied at every time step ($t = t + \Delta t$) to obtain the motions of individual atoms in a system. By incrementally increasing the time with a fixed time step, a series of atomic conformations is generated iteratively in time, which we designate a "system trajectory". As such, MD is a deterministic technique, meaning that Newton's equations are integrated reversibly in time. The direction in time is merely arbitrary; in theory it is possible to swap the sign of the atom velocities,

thereby going "back" in time [12]. Note that $\Delta t$ is determined by the fastest occurring process in the system. In the case of atomistic simulations, this is the vibration of hydrogen atom in range of one femtosecond ($10^{-15}$ s). The short time step is required to minimise inaccuracies in the cumulative integration of each force field equation.

There are several ways to integrate equation 2.9 numerically. One of the earliest methods is the velocity *Verlet* algorithm [46, 47]. In this method, a modified set of coordinates $r$ at a new time $(t + \Delta t)$ is calculated from a previous set of coordinates and accelerations at time $t$.

$$v(t + \frac{1}{2}\Delta t) = v(t) + \frac{\Delta t}{2m}F(t) \tag{2.12}$$

$$r(t + \Delta t) = r(t) + \Delta t v(t + \frac{1}{2}\Delta t) \tag{2.13}$$

In GROMACS [6], the standard integrator is the so-called *leap-frog* algorithm, which is basically a modification of the original *Verlet* algorithm. The integration steps are computed by:

$$v(t + \frac{1}{2}\Delta t) = v(t - \frac{1}{2}t) + \frac{\Delta t}{m}F(t) \tag{2.14}$$

$$r(t + \Delta t) = r(t) + \Delta t v(t + \frac{1}{2}\Delta t) \tag{2.15}$$

Hence, the velocities $v(t + \frac{1}{2}\Delta t)$ are based on the velocities at time $t - \frac{1}{2}\Delta t$, while the positions $r(t + \Delta t)$ are derived from the previous position $r(t)$ with intermediate velocity $v(t + \frac{1}{2}\Delta t)$. Or in other words: the velocities "skip" or *leap* over the positions during the integrations and vice versa [48].

It is essential that the integrator is as accurate as possible, as the system can become unstable if accumulated numerical errors propagate in time. Although both velocity Verlet and *leap-frog* can generate identical trajectories, the *leap-frog* integrator is considered to be computationally more stable.

A number of algorithms have been designed to speed up the MD simulation integration. The time step of 1 fs is increased to 2 fs by constraining the bond lengths in the system. That constraint "freezes" the bond lengths and angles after

integration of the forces in the system and is justified because temporary bond stretching vibrations are usually not coupled to global protein dynamics and function. In GROMACS, the LINCS algorithm (Linear Constraint Solver) is used as default to fix bond lengths and angles after the integration step, producing a four-time MD simulation speed up [49].

In ACEMD [50], another MD package used in this thesis, the M-SHAKE algorithm [51, 52] is applied for bond constraints and RATTLE for velocity constraints [53]. In addition, a hydrogen mass repartitioning scheme increases the time step to 4 fs, thereby increasing the integrator efficiency [50].

### 2.2.3 Periodic Boundary Conditions

As all particles interact with each other, there will obviously be artefacts at the boundaries of the simulation box. A "trick" called *periodic boundary conditions* (PBC) is therefore applied. PBC construct an infinite system by copying the simulation cell in all directions. By using this method, the system virtually has no boundaries, so e.g. a water molecule that exits the left side of the box reappears on the right side. As a result, the total amount of atoms (N) in the system remains constant in the main unit cell.

Different box types exist such as cubes, rhombic dodecahedrons and truncated octahedrons. Preferably, the size of the box should be as small as possible to limit the total amount of solvent atoms in the system, which consequently reduces the computational simulation cost. Importantly, the box size may not be too small, as periodic images can sense each other through long-range interactions. Typically, those interactions are much lower between 1.0 and 1.4 nm, it is therefore advised to create a box size with a minimum distance of 0.5 to 0.7 nm between each protein atom and the periodic box edges. The influence of the box shape on the dynamic properties of a protein in a MD simulation has been assessed previously [54].

### 2.2.4 Ensemble

An ensemble is a macroscopic collection consisting of different microscopic states of a system. Typically, a MD simulation is conducted using a chosen thermodynamic ensemble, i.e. a combination of a constant number of atoms (N), energy (E), pressure (P), volume (V) and temperature (T). Usually, a selection of three parameters defines the thermodynamic ensemble of a system. The most straightforward case is simply solving the equations mentioned above, which yields the microcanonical NVE ensemble with a fixed amount of particles (N), fixed volume

(V) and fixed total energy (E). That corresponds to an isolated system, as the total energy remains constant.

A better approximation of the experimental measurements is the canonical NVT ensemble, which applies an average constant temperature by using a specific algorithm called *thermostat*. That algorithm derives the temperature in the system from the atomic velocities, where each particle in the system is coupled to an external heat bath with a given temperature. The thermostat constantly adds or removes energy from the system by adjusting the velocities of each atom, which accordingly keeps the average temperature in the simulation box constant. For example, the Berendsen thermostat will generate a correct average temperature, but unfortunately it does not generate temperature distributions corresponding to a true statistical mechanical ensemble. As a result, the velocity rescaling thermostat is now commonly used in the GROMACS package, which accurately represents a canonical ensemble [55]. Other implementations are the Langevin [56] and Nose-Hoover thermostat [57, 58]. The differences between the thermostat algorithms are small: although temperature fluctuations are affected, the average temperature is almost identical in each case.

Alternatively, an isobaric-isothermal NPT ensemble can be applied using a *barostat* to keep the average pressure constant. Just like a simulation box is linked to a heat bath with a thermostat, so is there coupling to a pressure bath with a barostat. In contrast to NVT, NPT allows the volume to fluctuate by modifying the dimensions of the periodic box and coordinates in each time step, which creates an average reference pressure in the system. The NPT ensemble is essential to stabilise for example lipid bilayers with or without embedded membrane proteins and it is therefore frequently used to compare simulation data with experimentally measured properties.

Importantly, comparison with ensemble measured properties relies on the ergodic hypothesis, which states that all possible microstates can represent macroscopic properties if the complete phase space is sampled in the simulation trajectory. Hence, time averaged properties from MD simulation can be compared to experimental data. However, when sampling is insufficient, the system is considered nonergodic, meaning that only a subset of the phase space has been sampled. In a simulation trajectory, the system evolves into distinct microstates in time, but conformations can get trapped in local energy minima. Averaging over a number of snapshots in a simulation trajectory is therefore not representative for a thermodynamic property when conformational sampling is insufficient in time.

### 2.2.5 Solvent models

While we have now described the force field definitions of biomolecules, evidently the environment of immersed biomolecules needs to be characterised accurately as well. To this end, many different solvent models have been developed over the years to study the important effects of water molecules in binding pockets or the transport of water molecules through aquaporins [59]. The most commonly used water models are the explicit TIP3P and SPC, while more advanced models such as SPC/E, TIP4P and TIP5P provide higher accuracy (e.g. inclusion of the water dipole moment). In addition, implicit solvent models have been developed that use a continuum solvation description. By replacing the explicit representation of water molecules with an infinite continuum containing the dielectric properties of water, the amount of particles in a system is drastically reduced (see for example a study of the HIV protease by Hornak *et al.* [60]), thereby reducing computational costs and extending the accessible time scales. Solvation models such as Poisson-Boltzmann (PB) and generalised Born (GB) treat the solvent as a dielectric continuum, where mean interactions are modelled instead. We will explain these two methods more in detail in section 3.4.1.

### 2.2.6 Common protocol

While each simulation protocol will differ depending on the content of the system and the questions posed in advance, we will illustrate how to set up a MD simulation for a standard protein. Protein structure files called PDB files can be downloaded from the RCSB Protein Data Bank [61], where a large amount of structures determined by either X-ray crystallography, NMR or electron microscopy is available. In case of a crystal structure, hydrogens are only present in rare ultra resolution structures, so usually hydrogen atoms need to be added. In rare cases, all residues are present, but generally not all residues in the protein have been resolved or lack certain side chain atoms. Another issue is the determination of the correct protonation state of polar residues such as histidine [62, 63]. Also, glutamine and asparagine side chains often need to be flipped 180 degrees, as crystallography is not able to distinguish diffraction patterns from nitrogen and oxygen. In the cases where the protein structure has incomplete termini, it is advised to neutralise these ends by addition of an acetyl group (ACE) to the N-terminus or an N-methyl amide (NME) to the C-terminus. If the structure of a protein is unknown, it is possible to create a homology model if a suitable template structure can be found.

In summary, it is essential to verify the integrity of the protein structure before starting the simulation, as even small errors can propagate in time producing artefacts, thereby jeopardising the reliability of the simulation. MD is a deterministic technique, thus the output depends highly on the provided input structure. Also, even though for example geometric errors can be rather small, the force field is usually not able to "fix" this, as the structure might get trapped in local minima [64]. A notable example is the formation of a so-called "wedding ring", for example during the placement of residue side chains during homology modelling. As illustrated in Figure 2.5, the aromatic rings of hydrophobic protein residues are intertwined. Because bonds are unbreakable in molecular mechanics, the force field will never ever be able to correct this conformation and the starting structure simply has to be remodelled.

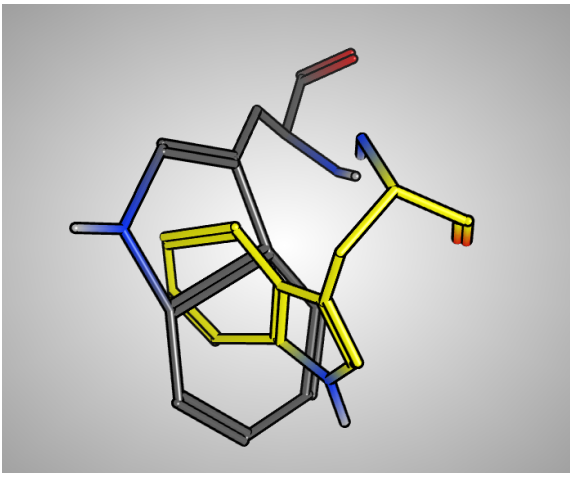

Figure 2.5: **Example of an unfortunate modelling error.** Two tryptophan residues adopt an intimate conformation due to incorrect placement of the side chain residues. Because bonds are unbreakable in a force field, a minimisation step is unable to solve this error.

After the selection of the starting structure, a molecular topology is defined from the force field parameters. Bonds, angles, dihedrals and atom pairs are contained in a parameter file, which cannot be changed during the course of the MD simulation. Standard force field parameters are available for common molecules in the system. However, for non-classical ligands or residues, parameters need to be determined by the user [65, 66] or need to be converted from other force fields, for example using the Acpype program [67].

After generation of the protein topology, the protein is placed in a virtual box with edges at a safe distance (usually around 1 nm) to avoid PBC artefacts (see section 2.2.1). The system is filled with solvent molecules and replacement of

random solvent molecules by counterions neutralises the charge of the protein. Alternatively, a predefined salt concentration can be specified to mimic experimental buffer conditions. Usually, the next step is to perform a short energy minimisation using steepest descent or conjugate gradient to resolve clashes in the system. Once the system potential energy has been minimised, an *equilibration run* is started. Initial velocities are assigned to all atoms in the system. A position restraint force is applied to the protein atoms to relax the solvent molecules, thereby filling any solvent-accessible cavities in the protein structure. Depending on the size and the stability of the protein, multiple minimisation and equilibration runs have to be performed. Once the potential energy of the system has reached a stable minimum, the restraints are released from the protein atoms and the so-called *production run* is initiated. The protein dynamics can now be followed in time. For decent sampling the simulation should be at least in the nanosecond range, depending on the size of the protein and the questions to be answered.

## 2.3 Limitations and challenges

MD simulations suffer from a number of restrictions, which are outlined more in detail below. The limitations can be summarised as: i) which representation of the system is appropriate to resemble experimental conditions, ii) the force field is inherently "imperfect", and iii) spatiotemporal restrictions hinder full sampling.

### 2.3.1 Levels of approximation

Every system that is studied in research is afflicted by certain inaccuracy. Evidently, the highest theoretical accuracy is preferred, but as with experimental studies, the cost of computational time has to be taken into account as well. Regarding MD simulations, a consensus should be found between cost and outcome of the simulations. To describe the motions of particles in a predefined system, one can resort to quantum physical methods, but imposing these calculations on macromolecular structures is not always feasible. In addition, depending on the posed questions, the outcome of those calculations would not necessarily result in a more accurate result. In other words, exploring the conformational landscape of a biomolecule can be compared to looking at a cartographic map, showing different tracks between the conformational states. Such a map can show every possible track, however, showing too much detail can be superfluous and confusing, as the main roads can be sufficient to reach a specific location.

With MD simulations, sometimes a more precise characterisation of the system is needed. For example, QM/MM systems treat a part of the system (usually a sphere, for example the binding site of a protein) by quantum mechanics while the remaining of the system (outside of the sphere) is treated with a faster molecular mechanics scheme. Considering the map analogy, the MM can be regarded as the overview of the map, while the QM part resembles an enlarged part of the map such as a city centre. The method can be used to study enzymatic reactions for example, where the active site is treated with QM to model bond formation and breaking, while the remaining of the protein is modelled by static-bonded MM.

Usually, most MD simulations are simulated using an all-atom representation, that is, all atoms are modelled explicitly as individual points, including all hydrogens atoms. There is a large number of force fields available using this representation, such as the AMBER, CHARMM and OPLS/AA force fields. In addition, so-called united-atom representations do not model apolar hydrogens explicitly. In that case, those apolar hydrogens are joined with carbon atoms to allow less atoms in the systems, thereby allowing faster calculations and thus longer accessible time scales. The GROMOS force fields are examples of this class, with parameters available for proteins, lipids and ribonucleotides. A course-grained representation reduces the amount of atoms in the system even further by fusing similar atoms together into beads as single interaction points. Course-graining is useful because the number of particles is decreased significantly in the simulation system. The degrees of freedom are reduced considerably and as a consequence the computational cost of the simulation is minimised. Though this simplification no longer allows tracking of individual atoms, course-graining can thus be beneficial when large complexes or processes at long time scales are of interest. For example, the Martini force field [68, 69] can be used on large protein complexes, such as modelling of the immature HIV virion [70]. In addition, the force field is frequently used for analysis of membrane systems, such as gating of ion channels [71], membrane pores [72], insertion of peptides [73, 74] and larger membrane proteins [75]. Sometimes, the system is first simulated using course-grained potentials, but is subsequently converted into an all-atom representation to study the biomolecular behaviour of the system more in detail [76–78]. A number of reviews discusses the applications of course grained MD simulations more in detail [79–81] and parameters are frequently being improved [82].

Finally, some words should be spent on the development of so-called polarisable force fields. All-atom force fields treat atoms as fixed particles, while in reality electron clouds are constantly deforming depending on their environment. Although computationally more expensive, the development of such polarisable force field is currently ongoing, including all-atom [83, 84] and course grained [85] optimised parameters.

### 2.3.2 Force field accuracy

The precision of molecular modelling simulations relies considerably on the quality of the underlying force field parameters that describe the total potential energy in a system. Mixing force fields from different roots is highly inaccurate, though there are some exceptions to this rule [86, 87]. Hence, the force field choice and selection of parameters found in the literature is one of the most important preparations for a computational chemist conducting MD simulations.

Due to the recent increase in computational power and development of advanced algorithms, the simulation of much longer time scales uncovered force field inaccuracies, such as force fields that are either "too $\alpha$-helical" [60, 88, 89] (i.e. overstabilise $\alpha$-helices in proteins compared to experiments) or contain dihedral parameter errors [64]. Hence, many force fields are optimised continuously by a meticulous refinement of the parameters through quantum mechanical calculations and careful comparison with experimental measurements. The force fields used in this thesis are OPLS/AA and AMBER ff99SB, which are both all-atom force fields with a similar functional form. The latter was obtained by adjustment of the backbone dihedral terms to reduce the overstabilisation of $\alpha$-helical conformations [60] and has been put forward as the force field of choice for simulations in the microsecond time scale [90]. The ff99SB force field has been further refined into AMBER ff99SB-ILDN, which contains adjustments for the side chain torsional parameters of isoleucine, leucine, aspartate and asparagine residues (hence the suffix "ILDN"). Those residues displayed significantly different rotameric distributions compared to statistics in the PDB and were accordingly corrected with new quantum mechanical calculations [91]. While this force field contains optimised parameters for side chain potentials, additional energy modifications of the backbone torsions were recently introduced to generate more balanced $\alpha$-helical propensities in MD simulations [92], resulting in the AMBER ff99SB*-ILDN force field [93]. In parallel, the CHARMM force field received similar backbone dihedral corrections, yielding CHARMM22* [93], and modifications to the side chain torsional parameters, designated CHARMM36 [94]. Similar corrections were recently made to the GROMOS parameter set as well [95].

Because there are many force fields available, which one should we select? Because each force field is inherently an approximation, the answer to this question depends of course on the scientific questions posed in advance. To this end, a number of systematic comparison studies have been performed to simplify answering that question. Some focus specifically on the folding [93, 96, 97] and denaturation [98] of proteins and/or use experimental validation by comparing with hydration free energies [99, 100] and with NMR spectroscopy data [90, 101, 102]. In addition, comparative force field studies have been performed for specific cases, such as the

reliability of nucleic acid force fields [103], the formation of hydrogen bonds [104], secondary structure tendencies [105–108] including residue specific propensities [109] and interactions between amino acid side chains [110]. Of course, simply investigating force field parameters can be misleading, as proteins are always immersed in a medium. Consequently, a number of noteworthy studies investigate the reliability of implicit [111–113] and explicit [114, 115] solvent models, protein-water interactions [116, 117] and lipid force fields [118–120]. In addition to force field parameters, the reader should also remember that algorithm implementations, electrostatic schemes like PME, box sizes and equilibration protocols can differ between all the above mentioned publications, which irrefutably impact experimental validation as well. Also, modelling failures are sometimes not caused by force field deficiencies, but rather by insufficient sampling of the protein conformational space. We will explain this "sampling" problem more in detail in the next section.

### 2.3.3 Sampling

MD simulations, just like experimental biophysical methods, are affected by spatiotemporal limitations. For experimental methods one tries to increase the resolution to investigate smaller or faster processes, while the resolution of MD simulations is already high, but the simulation of larger systems and/or longer processes requires a much higher computational cost. Hence, the limitations of MD simulations in space and time are reversed compared to experimental methods (as illustrated in Figure 2.2). Sampling problems in MD are related to the difficulty in spanning these large time scales. While force field inaccuracies are commonly held responsible, it has recently become more and more clear that inappropriate sampling can explain observed deviations from experiments [27].

How can we increase the sampling of biomolecules to longer time scales? Nowadays, a MD simulation of a small protein can be calculated for several nanoseconds on current desktop and laptop computers. But solving the equations of motions in a system is a complicated task, which involves solving the many different force field terms sequentially. The computational demand of MD simulations increases exponentially with the system size due to the pairwise non-bonded terms. As a result, sufficient conformational sampling of the energy landscape for larger complexes requires longer time scales. Choosing an appropriate representation as mentioned above can already reduce the spatiotemporal limitations. However, simulations in the range of microseconds and more and systems containing more than 100 000 atoms are currently only feasibly with high performance comput-

ing (HPC) infrastructure.[1] It is only recently that technological advances have increased the limit of biomolecular systems accessible by MD simulations, thereby allowing a broader overlap regarding the spatiotemporal resolution compared to experimental techniques. Clusters have been designed with unique chips and architecture optimised for MD simulations such as MD-GRAPE [121] and Anton [122]. Also, the advent of distributed computing such as folding@home [123, 124] and GPUGRID [125] allows the generation of many separate MD simulation trajectories.[2] Here, volunteers who have installed distributed computing software on their system dedicate idle CPU and/or GPU time to solve scientific questions in biochemistry, such as the folding of proteins or the binding of ligands to receptors. In additions, a large number of software developments such as parallelisation of the MD code and implementation of faster algorithms have increased the time limit accessible by MD simulations.

That of course raises the question: is the thermodynamic ensemble in the simulated system sampled sufficiently? For this, statistical analysis can be applied to ensure whether a thermodynamic ensemble is converged in time, i.e. if all relevant states have been probed adequately [126]. In addition, the construction of Markov state models (MSM) [127] allows a direct evaluation of the degree of sampling in a system by resolving measurable statistical properties from the simulation ensemble. It is expected that methods based on statistical relevance will become increasingly important in the field of biomolecular simulations.

# References

[1] Muybridge, E. (1887). Animal locomotion.
[2] Feynman, R. P. (1963). Feynman: Atoms in motion. In six easy pieces.
[3] Mulholland, A. J. (2008). Introduction. Biomolecular simulation. *J. R. Soc. Interface*. 5:S169–S172.
[4] Dror, R. O., Dirks, R. M., Grossman, J. P., Xu, H., and Shaw, D. E. (2012). Biomolecular Simulation: A Computational Microscope for Molecular Biology. *Annu. Rev. Biophys*. 41(1):429–452.
[5] Duerinckx, S. (2006). http://www.flickr.com/photos/darklife/91802770.
[6] Hess, B., Kutzner, C., van der Spoel, D., and Lindahl, E. (2008). GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput*. 4(3):435–447.
[7] Case, D. A., Cheatham, T. E., Darden, T., Gohlke, H., Luo, R., Merz, K. M., Onufriev, A., Simmerling, C., Wang, B., and Woods, R. J. (2005). The Amber biomolecular simulation programs. *J. Comput. Chem*. 26(16):1668–1688.

[1] We are grateful to the Flemish Supercomputer Centre. A considerable amount of simulations from chapter 4 and 6 have been conducted on their infrastructure.

[2] In chapter 5, we have applied GPUGRID to run multiple separate MD trajectories to explore the conformational ensemble of the HIV-1 gp41 FP. We therefore thank the GPUGRID volunteers who donated GPU computing time to the project.

[8] Brooks, B. R., Brooks, C. L., MacKerell, A. D., Nilsson, L., Petrella, R. J., Roux, B., Won, Y., Archontis, G., Bartels, C., and Boresch, S. (2009). CHARMM: the biomolecular simulation program. *J. Comput. Chem*. 30(10):1545–1614.

[9] Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R. D., Kale, L., and Schulten, K. (2005). Scalable molecular dynamics with NAMD. *J. Comput. Chem*. 26(16):1781–1802.

[10] Schlick, T., Collepardo-Guevara, R., Halvorsen, L. A., Jung, S., and Xiao, X. (2011). Biomolecular modeling and simulation: a field coming of age. *Q. Rev. Biophys*. 44(2):191–228.

[11] Hansson, T., Oostenbrink, C., and van Gunsteren, W. (2002). Molecular dynamics simulations. *Curr. Opin. Struct. Biol*. 12(2):190–196.

[12] Adcock, S. A. and McCammon, J. A. (2006). Molecular dynamics: survey of methods for simulating the activity of proteins. *Chem. Rev*. 106(5):1589–1615.

[13] Durrant, J. D. and McCammon, J. A. (2011). Molecular dynamics simulations and drug discovery. *BMC. Biol*. 9:71.

[14] van Gunsteren, W. F., Bakowies, D., Baron, R., Chandrasekhar, I., Christen, M., Daura, X., Gee, P., Geerke, D. P., Glättli, A., Hünenberger, P. H., Kastenholz, M. A., Oostenbrink, C., Schenk, M., Trzesniak, D., van der Vegt, N. F. A., and Yu, H. B. (2006). Biomolecular Modeling: Goals, Problems, Perspectives. *Angew. Chem. Int. Ed*. 45(25):4064–4092.

[15] Karplus, M. and Kuriyan, J. (2005). Molecular dynamics and protein function. *Proc. Natl. Acad. Sci. U. S. A*. 102(19):6679–6685.

[16] Guliaev, A. B., Cheng, S., and Hang, B. (2012). Protein dynamics via computational microscope. *World. J. Methodol*. 2(6):42–49.

[17] Morra, G., Meli, M., and Colombo, G. (2008). Molecular dynamics simulations of proteins and peptides: From folding to drug design. *Curr. Protein Pept. Sci*. 9(2):181–196.

[18] Salsbury Jr, F. R. (2010). Molecular dynamics simulations of protein dynamics and their relevance to drug discovery. *Curr. Opin. Pharmacol*. 10(6):738–744.

[19] Chodera, J. D., Mobley, D. L., Shirts, M. R., Dixon, R. W., Branson, K., and Pande, V. S. (2011). Alchemical free energy methods for drug discovery: progress and challenges. *Curr. Opin. Struct. Biol*. 21(2):150–160.

[20] Harvey, M. J. and De Fabritiis, G. (2012). High-throughput molecular dynamics: the powerful new tool for drug discovery. *Drug Discov. Today*. 17(19-20):1059–1062.

[21] Sinko, W., Lindert, S., and McCammon, J. A. (2013). Accounting for receptor flexibility and enhanced sampling methods in computer-aided drug design. *Chem. Biol. Drug Des*. 81(1):41–49.

[22] Alonso, H., Bliznyuk, A. A., and Gready, J. E. (2006). Combining docking and molecular dynamic simulations in drug design. *Med. Res. Rev*. 26(5):531–568.

[23] Best, R. B. (2012). Atomistic molecular simulations of protein folding. *Curr. Opin. Struct. Biol*. 22(1):52–61.

[24] Dill, K. A. and MacCallum, J. L. (2012). The Protein-Folding Problem, 50 Years On. *Science*. 338(6110):1042–1046.

[25] Freddolino, P. L., Harrison, C. B., Liu, Y., and Schulten, K. (2010). Challenges in protein-folding simulations. *Nat. Phys*. 6(10):751–758.

[26] Snow, C. D., Sorin, E. J., Rhee, Y. M., and Pande, V. S. (2005). How well can simulation predict protein folding kinetics and thermodynamics. *Annu. Rev. Biophys*. 34(1):43–69.

[27] Mobley, D. L. (2011). Let's get honest about sampling. *J. Comput. Aided Mol. Des*. 26(1):93–95.

[28] Borhani, D. W. and Shaw, D. E. (2011). The future of molecular dynamics simulations in drug discovery. *J. Comput. Aided Mol. Des*. 26(1):15–26.

[29] Cui, Y. (2010). Using molecular simulations to probe pharmaceutical materials. *J. Pharm. Sci*. 100(6):2000–2019.

[30] Zwier, M. C. and Chong, L. T. (2010). Reaching biological timescales with all-atom molecular dynamics simulations. *Curr. Opin. Pharmacol*. 10(6):745–752.

[31] Klepeis, J. L., Lindorff-Larsen, K., Dror, R. O., and Shaw, D. E. (2009). Long-timescale molecular dynamics simulations of protein structure and function. *Curr. Opin. Struct. Biol.* 19(2):120–127.
[32] Lane, T. J., Shukla, D., Beauchamp, K. A., and Pande, V. S. (2013). To milliseconds and beyond: challenges in the simulation of protein folding. *Curr. Opin. Struct. Biol.* 23(1):58–65.
[33] Fuentes, G., Dastidar, S. G., Madhumalar, A., and Verma, C. S. (2010). Role of protein flexibility in the discovery of new drugs. *Drug Dev. Res.* 72(1):26–35.
[34] Marsh, J. A., Teichmann, S. A., and Forman-Kay, J. D. (2012). Probing the diverse landscape of protein flexibility and binding. *Curr. Opin. Struct. Biol.* 22(5):643–650.
[35] Schneider, G. (2010). Virtual screening: an endless staircase? *Nat. Res. Drug. Discov.* 9(4):273–276.
[36] Okimoto, N., Futatsugi, N., Fuji, H., Suenaga, A., Morimoto, G., Yanai, R., Ohno, Y., Narumi, T., and Taiji, M. (2009). High-performance drug discovery: computational screening by combining docking and molecular dynamics simulations. *PLoS Comput. Biol.* 5(10):e1000528.
[37] Nichols, S. E., Baron, R., Ivetac, A., and McCammon, J. A. (2011). Predictive power of molecular dynamics receptor structures in virtual screening. *J. Chem. Inf. Model.* 51(6):1439–1446.
[38] Osguthorpe, D. J., Sherman, W., and Hagler, A. T. (2012). Exploring Protein Flexibility: Incorporating Structural Ensembles From Crystal Structures and Simulation into Virtual Screening Protocols. *J. Phys. Chem. B.* 116(23):6952–6959.
[39] Rastelli, G. (2013). Emerging topics in structure-based virtual screening. *Pharm. Res.* 30(5):1458–1463.
[40] Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W., and Kollman, P. A. (1995). A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* 117(19):5179–5197.
[41] Robertson, A., Luttmann, E., and Pande, V. S. (2008). Effects of long-range electrostatic forces on simulated protein folding kinetics. *J. Comput. Chem.* 29(5):694–700.
[42] Darden, T., York, D., and Pedersen, L. (1993). Particle mesh Ewald: An Nlog(N) method for Ewald sums in large systems. *J. Chem. Phys.* 98(12):10089–10092.
[43] Essmann, U., Perera, L., Berkowitz, M. L., Darden, T., Lee, H., and Pedersen, L. G. (1995). A smooth particle mesh Ewald method. *J. Chem. Phys.* 103(19):8577–8593.
[44] Hess, B., van der Spoel, D., and Lindahl, E. (2010). Gromacs User Manual Version 4.5. University of Groningen, Netherland.
[45] Deckers, S. M., Venken, T., Khalesi, M., Gebruers, K., Baggerman, G., Lorgouilloux, Y., Shokribousjein, Z., Ilberg, V., Schonberger, C., and Titze, J. (2012). Combined Modeling and Biophysical Characterisation of CO2 Interaction with Class II Hydrophobins: New Insight into the Mechanism Underpinning Primary Gushing. *J. Am. Soc. Brew. Chem.* 70(4):249–256.
[46] Verlet, L. and Levesque, D. (1967). On the theory of classical fluids VI. *Physica.* 36(2):254–268.
[47] Swope, W. C. (1982). A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *J. Chem. Phys.* 76(1):637–649.
[48] Allen, M. P. and Tildesley, D. J. (1989). Computer Simulation of Liquids. Oxford Univ Press.
[49] Hess, B., Bekker, H., Berendsen, H., and Fraaije, J. (1997). LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.* 18(12):1463–1472.
[50] Harvey, M. J., Giupponi, G., and Fabritiis, G. D. (2009). ACEMD: Accelerating Biomolecular Dynamics in the Microsecond Time Scale. *J. Chem. Theory Comput.* 5(6):1632–1639.
[51] Ryckaert, J.-P., Ciccotti, G., and Berendsen, H. J. C. (1977). Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* 23(3):327–341.
[52] Kräutler, V., van Gunsteren, W. F., and Hünenberger, P. H. (2001). A fast SHAKE algorithm to solve distance constraint equations for small molecules in molecular dynamics simulations. *J. Comput. Chem.* 22(5):501–508.
[53] Andersen, H. C. (1983). Rattle: A "velocity" version of the shake algorithm for molecular dynamics calculations. *J. Comput. Phys.* 52(1):24–34.

[54] Wassenaar, T. A. and Mark, A. E. (2005). The effect of box shape on the dynamic properties of proteins simulated under periodic boundary conditions. *J. Comput. Chem*. 27(3):316–325.
[55] Bussi, G., Donadio, D., and Parrinello, M. (2007). Canonical sampling through velocity rescaling. *J. Chem. Phys*. 126(1):014101.
[56] Schneider, T. and Stoll, E. (1978). Molecular-dynamics study of a three-dimensional one-component model for distortive phase transitions. *Phys. Rev. B*. 17(3):1302.
[57] Nosé, S. (1984). A molecular dynamics method for simulations in the canonical ensemble. *Mol. Phys*. 52(2):255–268.
[58] Hoover, W. G. (1985). Canonical dynamics: Equilibrium phase-space distributions. *Phys. Rev. A*. 31(3):1695.
[59] de Groot, B. L. (2001). Water Permeation Across Biological Membranes: Mechanism and Dynamics of Aquaporin-1 and GlpF. *Science*. 294(5550):2353–2357.
[60] Hornak, V., Okur, A., Rizzo, R., and Simmerling, C. (2006). HIV-1 protease flaps spontaneously open and reclose in molecular dynamics simulations. *Proc. Natl. Acad. Sci. U. S. A*. 103(4):915–920.
[61] Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., and Tasumi, M. (1977). The Protein Data Bank. A computer-based archival file for macromolecular structures. *Eur. J. Biochem*. 80(2):319–324.
[62] Ben-Shimon, A., Shalev, D. E., and Niv, M. Y. (2013). Protonation States in Molecular Dynamics Simulations of Peptide Folding and Binding. *Curr. Pharm. Des*. 19(23):4173–4181.
[63] Onufriev, A. V. and Alexov, E. (2013). Protonation and pK changes in protein–ligand binding. *Q. Rev. Biophys*. 46(02):181–209.
[64] Schreiner, E., Trabuco, L. G., Freddolino, P. L., and Schulten, K. (2011). Stereochemical errors and their implications for molecular dynamics simulations. *BMC Bioinformatics*. 12(1):190.
[65] Vanquelef, E., Simon, S., Marquant, G., Garcia, E., Klimerak, G., Delepine, J. C., Cieplak, P., and Dupradeau, F.-Y. (2011). RED Server: a web service for deriving RESP and ESP charges and building force field libraries for new molecules and molecular fragments. *Nucleic Acids Res*. 39:W511–W517.
[66] Lemkul, J. A., Allen, W. J., and Bevan, D. R. (2010). Practical considerations for building GROMOS-compatible small-molecule topologies. *J. Chem. Inf. Model*. 50(12):2221–2235.
[67] Sousa da Silva, A. W. and Vranken, W. F. (2012). ACPYPE - AnteChamber PYthon Parser interfacE. *BMC Res. Notes*. 5(1):367.
[68] Marrink, S. J., Risselada, H. J., Yefimov, S., Tieleman, D. P., and De Vries, A. H. (2007). The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations. *J. Phys. Chem. B*. 111(27):7812–7824.
[69] Monticelli, L., Kandasamy, S. K., Periole, X., Larson, R. G., Tieleman, D. P., and Marrink, S.-J. (2008). The MARTINI Coarse-Grained Force Field: Extension to Proteins. *J. Chem. Theory Comput*. 4(5):819–834.
[70] Bond, P. J., Wee, C. L., and Sansom, M. S. P. (2008). Coarse-Grained Molecular Dynamics Simulations of the Energetics of Helix Insertion into a Lipid Bilayer. *Biochemistry*. 47(43):11321–11331.
[71] Dryga, A., Chakrabarty, S., Vicatos, S., and Warshel, A. (2012). Realistic simulation of the activation of voltage-gated ion channels. *Proc. Natl. Acad. Sci. U. S. A*. 109(9):3335–3340.
[72] Marrink, S. J., De Vries, A. H., and Tieleman, D. P. (2009). Lipids on the move: simulations of membrane pores, domains, stalks and curves. *Biochim. Biophys. Acta*. 1788(1):149–168.
[73] Gkeka, P. and Sarkisov, L. (2010). Interactions of Phospholipid Bilayers with Several Classes of Amphiphilic $\alpha$-Helical Peptides: Insights from Coarse-Grained Molecular Dynamics Simulations. *J. Phys. Chem. B*. 114(2):826–839.
[74] Hwang, H., Schatz, G. C., and Ratner, M. A. (2009). Coarse-Grained Molecular Dynamics Study of Cyclic Peptide Nanotube Insertion into a Lipid Bilayer. *J. Phys. Chem. A*. 113(16):4780–4787.
[75] Bond, P. J. and Sansom, M. S. P. (2006). Insertion and Assembly of Membrane Proteins via Simulation. *J. Am. Chem. Soc*. 128(8):2697–2704.

[76] Thøgersen, L., Schiøtt, B., Vosegaard, T., Nielsen, N. C., and Tajkhorshid, E. (2008). Peptide Aggregation and Pore Formation in a Lipid Bilayer: A Combined Coarse-Grained and All Atom Molecular Dynamics Study. *Biophys. J.* 95(9):4337–4347.

[77] Parton, D. L., Akhmatskaya, E. V., and Sansom, M. S. P. (2012). Multiscale Simulations of the Antimicrobial Peptide Maculatin 1.1: Water Permeation through Disordered Aggregates. *J. Phys. Chem. B*. 116(29):8485–8493.

[78] Stansfeld, P. J. and Sansom, M. S. P. (2011). From Coarse Grained to Atomistic: A Serial Multiscale Approach to Membrane Protein Simulations. *J. Chem. Theory Comput*. 7(4):1157–1166.

[79] Rader, A. J. (2010). Coarse-grained models: getting more with less. *Curr. Opin. Pharmacol*. 10(6):753–759.

[80] Takada, S. (2012). Coarse-grained molecular simulations of large biomolecules. *Curr. Opin. Struct. Biol*. 22(2):130–137.

[81] Noid, W. G. (2013). Perspective: Coarse-grained models for biomolecular systems. *J. Chem. Phys*. 139(9):090901.

[82] De Jong, D. H., Singh, G., Bennett, W. F. D., Arnarez, C., Wassenaar, T. A., Schäfer, L. V., Periole, X., Tieleman, D. P., and Marrink, S. J. (2013). Improved Parameters for the Martini Coarse-Grained Protein Force Field. *J. Chem. Theory Comput*. 9(1):687–697.

[83] Ponder, J. W., Wu, C., Ren, P., Pande, V. S., Chodera, J. D., Schnieders, M. J., Haque, I., Mobley, D. L., Lambrecht, D. S., DiStasio Jr., R. A., Head-Gordon, M., Clark, G. N. I., Johnson, M. E., and Head-Gordon, T. (2010). Current Status of the AMOEBA Polarizable Force Field. *J. Phys. Chem. B*. 114(8):2549–2564.

[84] Cieplak, P., Dupradeau, F.-Y., Duan, Y., and Wang, J. (2009). Polarization effects in molecular mechanical force fields. *J. Phys.: Condens. Matter*. 21(33):333102.

[85] Yesylevskyy, S. O., Schäfer, L. V., Sengupta, D., and Marrink, S. J. (2010). Polarizable Water Model for the Coarse-Grained MARTINI Force Field. *PLoS Comput. Biol*. 6(6):e1000810.

[86] Sapay, N. and Tieleman, D. P. (2010). Combination of the CHARMM27 force field with united-atom lipid force fields. *J. Comput. Chem*. 32(7):1400–1410.

[87] Cordomí, A., Caltabiano, G., and Pardo, L. (2012). Membrane Protein Simulations Using AMBER Force Field and Berger Lipid Parameters. *J. Chem. Theory Comput*. 8(3):948–958.

[88] Yoda, T., Sugita, Y., and Okamoto, Y. (2004). Comparisons of force fields for proteins by generalized-ensemble simulations. *Chem. Phys. Lett*. 386(4-6):460–467.

[89] Best, R. B., Buchete, N.-V., and Hummer, G. (2008). Are Current Molecular Dynamics Force Fields too Helical? *Biophys. J*. 95(1):L07–L09.

[90] Lange, O. F., van der Spoel, D., and de Groot, B. L. (2010). Scrutinizing Molecular Mechanics Force Fields on the Submicrosecond Timescale with NMR Data. *Biophys. J*. 99(2):647–655.

[91] Lindorff-Larsen, K., Piana, S., Palmo, K., Maragakis, P., Klepeis, J. L., Dror, R. O., and Shaw, D. E. (2010). Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins*. 78(8):1950–1958.

[92] Best, R. B. and Hummer, G. (2009). Optimized Molecular Dynamics Force Fields Applied to the HelixCoil Transition of Polypeptides. *J. Phys. Chem. B*. 113(26):9004–9015.

[93] Piana, S., Lindorff-Larsen, K., and Shaw, D. E. (2011). How Robust Are Protein Folding Simulations with Respect to Force Field Parameterization? *Biophys J*. 100(9):L47–L49.

[94] Best, R. B., Zhu, X., Shim, J., Lopes, P. E. M., Mittal, J., Feig, M., and MacKerell, A. D., Jr. (2012). Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone $\phi$, $\psi$ and Side-Chain $\chi_1$ and $\chi_2$ Dihedral Angles. *J. Chem. Theory Comput*. 8(9):3257–3273.

[95] Schmid, N., Eichenberger, A. P., Choutko, A., Riniker, S., Winger, M., Mark, A. E., and Gunsteren, W. F. (2011). Definition and testing of the GROMOS force-field versions 54A7 and 54B7. *Eur. Biophys. J*. 40(7):843–856.

[96] Freddolino, P. L., Park, S., Roux, B., and Schulten, K. (2009). Force Field Bias in Protein Folding Simulations. *Biophys J*. 96(9):3772–3780.

[97] Best, R. B. and Mittal, J. (2010). Balance between $\alpha$ and $\beta$ Structures in Ab Initio Protein Folding. *J. Phys. Chem. B*. 114(26):8790–8798.

[98] Baxa, M. C., Haddadian, E. J., Jha, A. K., Freed, K. F., and Sosnick, T. R. (2012). Context and Force Field Dependence of the Loss of Protein Backbone Entropy upon Folding Using Realistic Denatured and Native State Ensembles. *J. Am. Chem. Soc*. 134(38):15929–15936.

[99] Shirts, M. R., Pitera, J. W., Swope, W. C., and Pande, V. S. (2003). Extremely precise free energy calculations of amino acid side chain analogs: Comparison of common molecular mechanics force fields for proteins. *J. Chem. Phys*. 119(11):5740–5761.

[100] Mobley, D. L., Dumont, É., Chodera, J. D., and Dill, K. A. (2010). Comparison of charge models for fixed-charge force fields: small molecule hydration free energies in explicit solvent. *J. Phys. Chem. B*. 115(5):1329–1332.

[101] Lindorff-Larsen, K., Maragakis, P., Piana, S., Eastwood, M. P., Dror, R. O., and Shaw, D. E. (2012). Systematic Validation of Protein Force Fields against Experimental Data. *PLoS One*. 7(2):e32131.

[102] Beauchamp, K. A., Lin, Y.-S., Das, R., and Pande, V. S. (2012). Are Protein Force Fields Getting Better? A Systematic Benchmark on 524 Diverse NMR Measurements. *J. Chem. Theory Comput*. 8(4):1409–1414.

[103] Reddy, S. Y., Leclerc, F., and Karplus, M. (2003). DNA Polymorphism: A Comparison of Force Fields for Nucleic Acids. *Biophys J*. 84(3):1421–1449.

[104] Paton, R. S. and Goodman, J. M. (2009). Hydrogen Bonding and $\pi$-Stacking: How Reliable are Force Fields? A Critical Evaluation of Force Field Descriptions of Nonbonded Interactions. *J. Chem. Inf. Model*. 49(4):944–955.

[105] Cino, E. A., Choy, W.-Y., and Karttunen, M. (2012). Comparison of Secondary Structure Formation Using 10 Different Force Fields in Microsecond Molecular Dynamics Simulations. *J. Chem. Theory Comput*. 8(8):2725–2740.

[106] Matthes, D. and de Groot, B. L. (2009). Secondary Structure Propensities in Peptide Folding Simulations: A Systematic Comparison of Molecular Mechanics Interaction Schemes. *Biophys J*. 97(2):599–608.

[107] Vymětal, J. and Vondrášek, J. (2013). Critical Assessment of Current Force Fields. Short Peptide Test Case. *J. Chem. Theory Comput*. 9(1):441–451.

[108] Gerben, S. R., Lemkul, J. A., Brown, A. M., and Bevan, D. R. (2013). Comparing atomistic molecular mechanics force fields for a difficult target: a case study on the Alzheimer's amyloid $\beta$-peptide. *J. Biomol. Struct. Dyn*. http://dx.doi.org/10.1080/07391102.2013.838518.

[109] Best, R. B., De Sancho, D., and Mittal, J. (2012). Residue-Specific & $\alpha$-Helix Propensities from Molecular Simulation. *Biophys. J*. 102(6):1462–1467.

[110] De Jong, D. H., Periole, X., and Marrink, S. J. (2012). Dimerization of Amino Acid Side Chains: Lessons from the Comparison of Different Force Fields. *J. Chem. Theory Comput*. 8(3):1003–1014.

[111] Zhang, L. Y., Gallicchio, E., Friesner, R. A., and Levy, R. M. (2001). Solvent models for protein–ligand binding: Comparison of implicit solvent Poisson and surface generalized Born models with explicit solvent simulations. *J. Comput. Chem*. 22(6):591–607.

[112] Roe, D. R., Okur, A., Wickstrom, L., Hornak, V., and Simmerling, C. (2007). Secondary Structure Bias in Generalized Born Solvent Models: Comparison of Conformational Ensembles and Free Energy of Solvent Polarization from Explicit and Implicit Solvation. *J. Phys. Chem. B*. 111(7):1846–1857.

[113] Juneja, A., Ito, M., and Nilsson, L. (2013). Implicit Solvent Models and Stabilizing Effects of Mutations and Ligands on the Unfolding of the Amyloid $\beta$-Peptide Central Helix. *J. Chem. Theory Comput*. 9(1):834–846.

[114] Paschek, D., Day, R., and Garcia, A. E. (2011). Influence of water–protein hydrogen bonding on the stability of Trp-cage miniprotein. A comparison between the TIP3P and TIP4P-Ew water models. *Phys. Chem. Chem. Phys*. 13(44):19840.

[115] Zielkiewicz, J. (2005). Structural properties of water: Comparison of the SPC, SPCE, TIP4P, and TIP5P models of water. *J. Chem. Phys*. 123(10):104501.

[116] Nerenberg, P. S. and Head-Gordon, T. (2011). Optimizing Protein-Solvent Force Fields to Reproduce Intrinsic Conformational Preferences of Model Peptides. *J. Chem. Theory Comput*. 7(4):1220–1230.

[117] Nerenberg, P. S., Jo, B., So, C., Tripathy, A., and Head-Gordon, T. (2012). Optimizing Solute-Water van der Waals Interactions To Reproduce Solvation Free Energies. *J. Phys. Chem. B*. 116(15):4524–4534.

[118] Piggot, T. J., Piñeiro, Á., and Khalid, S. (2012). Molecular Dynamics Simulations of Phosphatidylcholine Membranes: A Comparative Force Field Study. *J. Chem. Theory Comput*. 8(11):4593–4609.

[119] Siu, S. W. I., Vácha, R., Jungwirth, P., and Böckmann, R. A. (2008). Biomolecular simulations of membranes: Physical properties from different force fields. *J. Chem. Phys*. 128(12):125103.

[120] Jämbeck, J. P. M., Mocci, F., Lyubartsev, A. P., and Laaksonen, A. (2013). Partial atomic charges and their impact on the free energy of solvation. *J. Comput. Chem*. 34(3):187–197.

[121] Komeiji, Y., Uebayasi, M., Takata, R., Shimizu, A., Itsukashi, K., and Taiji, M. (1997). Fast and accurate molecular dynamics simulation of a protein using a specialpurpose computer. *J. Comput. Chem*. 18(12):1546–1563.

[122] Shaw, D. E., Dror, R. O., Salmon, J. K., Grossman, J. P., Mackenzie, K. M., Bank, J. A., Young, C., Deneroff, M. M., Batson, B., and Bowers, K. J. (2009). Millisecond-scale molecular dynamics simulations on Anton. *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis*. 39.

[123] Larson, S. M., Snow, C. D., Shirts, M., and Pande, V. S. (2009). Folding@ Home and Genome@ Home: Using distributed computing to tackle previously intractable problems in computational biology. *arXiv:09010866*.

[124] Beberg, A. L., Ensign, D. L., Jayachandran, G., Khaliq, S., and Pande, V. S. (2009). Folding@ home: Lessons from eight years of volunteer distributed computing. *Parallel & Distributed Processing, IEEE International Symposium*. 1–8.

[125] Buch, I., Harvey, M. J., Giorgino, T., Anderson, D. P., and De Fabritiis, G. (2010). High-Throughput All-Atom Molecular Dynamics Simulations Using Distributed Computing. *J. Chem. Inf. Model*. 50(3):397–403.

[126] Shirts, M. R. (2013). Simple Quantitative Tests to Validate Sampling from Thermodynamic Ensembles. *J. Chem. Theory Comput*. 9(2):909–926.

[127] Pande, V. S., Beauchamp, K., and Bowman, G. R. (2010). Everything you wanted to know about Markov State Models but were afraid to ask. *Methods*. 52(1):99–105.

# Chapter 3

## Binding free energy calculations

"In An Isolated System; The Entropy Can Only Increase"

The 2nd Law - Muse

## 3.1 Introduction

Without thorough analysis, a MD simulation is merely a suggestive and expensive movie. We will illustrate an important analysis applied to MD used throughout this thesis: the quantitative estimation of the binding free energy between two interacting particles. Investigation of the binding between molecules is one of the most fundamental aspects in biophysical research, as molecular recognition is essential in practically all biological processes. Extracting binding free energies from molecular simulations is thus potentially interesting to understand underlying processes at the atomic scale. The affinity of biological systems can be estimated when thermodynamic measurements in experiments are unfeasible. In addition, measuring binding free energy values with an appropriate precision could significantly reduce the need of those wet lab experiments, which are often expensive, difficult and time-consuming.

For these reasons, there is a lot of interest from the pharmaceutical industry to accurately predict biological activities using *in silico* methods. Because drug development is a highly time-absorbing task (requires usually more than 10 years) and expensive (costs approach one billion dollar per drug on the market), computer-aided drug design (CADD) can be applied to speed up the process. For example, binding free energies of potential drugs in interaction with the target of interest can be evaluated. Previously, traditional screening of drug molecules did not rely on any protein target information. Currently however, pharmaceutical companies are incorporating rational design strategies in their drug discovery pipelines to complement or even replace traditional screening techniques. To this end, it is crucial

to obtain trustworthy predictive models of the binding affinity between potential hit molecules and their respective targets, to discard low potency compounds and inquire optimisation strategies.

Typically, computational methods are used in multiple stages of the drug discovery pipeline. In the early stages, virtual screening methods can be applied to identify initial "hits", that is, a small molecule that binds effectively to the target of interest and is a good candidate to optimise further into an effective drug by modifying molecular properties. In these stages, extensive virtual molecular databases can be screened, which requires fast computational filtering techniques. The time spent to evaluate a single compound must be short but may affect the reliability of the applied scoring function. However, once interesting hits have been found, designated as "lead" molecules, the accurate prediction of binding affinities becomes more important than speed. To optimise the initial hits and its properties using sequential chemical remodelling, effective ranking of good and weak affinity molecules is needed. These affinity scores can subsequently be correlated with the structural characteristics of the molecules, for example in structure-activity relationship (SAR) studies.

The first computational binding free energy methods were developed in the late 1980s and initially resulted in high expectations. It was envisioned that researchers in the near future would be able to simply push a button on a computer, which swiftly and accurately would offer a solution of a putative drug. In retrospect, these exceptions were probably way too high, as the high number of failures surpassed the number of early successes. Looking back, it is clear that significant hurdles had to be surmounted due to limitations at that time. Despite recent progress, there are still many question and challenges to be solved. Many researcher are now developing and optimising algorithms and tools to accurately describe non-covalent interactions by enhancing sampling techniques and utilising recent progress in computer hardware.

In the ideal case, the computationally calculated binding affinities would deviate from the experimental measurements by just 1 to 2 kcal $\text{mol}^{-1}$ [1]. Although an accurate estimation of binding affinity has been considered a "holy grail" in computational chemistry for many years [2, 3], we are still far from achieving this goal with the current set of algorithms. Binding free energy calculations are challenging due to the large number of interactions that need to be computed. Moreover, chemical effects such as protonation state and tautomeric conformations can be important. Other issues are the many degrees of freedom of the molecules involved, the reduction of conformational entropy upon binding, and the approximations present in the applied force field and solvent models. Consequently, a binding affinity analysis of a range of compounds should always take these

challenges and limitations into account, as frequently one of these gives rise to deviations from experimental affinities. An incorrect ranking of drug molecules can subsequently result in misguided and unsuccessful drug design strategies.

Generally, three different approaches can be distinguished. First, **exact methods** or so-called "alchemical" methods apply statistical mechanics to compute binding free energy differences between bound and unbound states of a receptor/ligand pair using a non-physical pathway (hence the term "alchemical"). These methods are considered the most rigorous and accurate to estimate relative binding affinities. Unfortunately, the calculations require a considerable amount of conformational sampling to obtain thermodynamically converging results, especially on larger protein/ligand complexes. That makes exact methods computationally consuming and is worsened even more when highly dissimilar ligands are compared. Examples of that class are thermodynamic integration (TI), free energy perturbation (FEP) and Bennet's acceptance ratio (BAR) [3–6]. Second, so-called **pathway methods** determine absolute binding free energy values by displacement of a ligand along a reaction coordinate, generating a potential of mean force (PMF). These pathways can be generated using steered MD (SMD) [7], umbrella sampling [8–10] or metadynamics [11]. Third, so-called **endpoint methods** are considered more approximate and only require simulations of the bound and unbound state of the ligands. Hence, the latter technique is more efficient for larger molecules such as peptide binding or scrutiny of protein-protein interactions. Molecular mechanics/Poisson-Boltzmann surface area (MM-PBSA), molecular mechanics/generalised-Born surface area (MM-GBSA) and linear interaction energy (LIE) are examples of that class [5, 6, 12, 13].

Due to the relatively large size of the ligands and proteins investigated in this manuscript, exact and pathway methods are outside the scope of this thesis and are discussed more in detail in the following review articles [1, 14–16]. A number of articles discusses the implementation, application and differences between those binding free energy methods [4, 17, 18]. Below, we will give a summary of the theory behind non-covalent binding and the approximate binding free energy methods used in this thesis, including an emphasis of the applications and current limitations.

## 3.2 Theory of non-covalent interactions

### 3.2.1 Kinetics

Considering two molecules in solution, the reversible binding of both molecules can be expressed as a chemical reaction by:

$$A + B \underset{k_{off}}{\overset{k_{on}}{\rightleftharpoons}} AB \tag{3.1}$$

Here, $A$ is the first free molecule, $B$ is the second free molecule and $AB$ is the complex of both molecules. These two molecules can be either kind, for example two proteins of similar size, a protein-peptide interaction or the binding of a ligand into a protein cavity. The kinetics can be expressed using a first-order model with $k_{on}$ and $k_{off}$ the rate constant for respectively association of $A$ and $B$ and dissociation of the $AB$ complex. Upon equilibrium, a combination of $A$, $B$ and $AB$ will exist in the system. The equation can be written using the equilibrium association constant $K_a$, or conversely the equilibrium dissociation constant $K_d$:

$$K_a = \frac{1}{K_d} = \frac{k_{on}}{k_{off}} = \frac{[AB]_{eq}}{[A]_{eq}[B]_{eq}} \tag{3.2}$$

where $[AB]_{eq}$, $[A]_{eq}$ and $[B]_{eq}$ are the concentrations at equilibrium of the complex $AB$ and the constituents $A$ and $B$. Higher values of $K_a$ thus indicate a larger attraction between both molecules. Consequently, $K_a$ is a measure of the binding affinity. However, usually $K_d$ is applied to compare binding affinity values: a compound with nanomolar affinity ($10^{-9}$ M) is therefore a stronger binder than a compound with only micromolar affinity ($10^{-6}$ M).

### 3.2.2 Thermodynamics

The association constant $K_a$ can be expressed thermodynamically in terms of the concentration-independent Gibbs free energy of binding in standard experimental conditions (i.e. room temperature and 1 atmospheric pressure in a NPT ensemble) using the *van't Hoff* equation:

$$\Delta G^{\circ}_{bind} = -RTlnK_a \tag{3.3}$$

Here, R is the ideal gas constant and T is the temperature. For clarity, each $\Delta G_{bind}$ in this manuscript refers to a binding free energy in standard conditions $\Delta G^{\circ}_{bind}$. Relating the Gibbs free energy with the equilibrium constant shows that higher $K_a$ values, and hence higher concentrations of complex at equilibrium, result in more negative $\Delta G$ values. A reaction can occur spontaneously with a negative $\Delta G$ (known as an exergonic reaction), while reactions with a positive $\Delta G$ value do not proceed spontaneously (known as an endergonic reaction) and require coupling to a favourable reaction. Or to put in other words: if the free energy of the complex $AB$ is smaller than the sum of its constituents ($A$ and $B$), then the reaction will occur spontaneously until an equilibrium is reached:

$$\Delta G^{\circ}_{bind} = G^{\circ}_{AB} - G^{\circ}_{A} - G^{\circ}_{B} \tag{3.4}$$

While the above equation results in an **absolute binding free energy difference** for one complex ($\Delta G$), it is typical to compute the **relative binding free energy difference** between different complexes ($\Delta\Delta G$). For example, ligands can attach to a receptor in a wild type and a mutated form. In that case, the relative binding free energy difference is calculated from the absolute binding free energies of the wild type and the ligand ($\Delta G^{\circ}_{bind1}$) and mutated receptors and the ligand ($\Delta G^{\circ}_{bind2}$) (as illustrated in Figure 3.1). Conversely, the receptor conformation can remain similar to compare the affinity of different ligands.

$$\Delta\Delta G^{\circ}_{bind} = \Delta G^{\circ}_{bind1} - \Delta G^{\circ}_{bind2} \tag{3.5}$$

$$= \Delta G^{\circ}_{wt} - \Delta G^{\circ}_{mutant} \tag{3.6}$$

Binding reactions have kcal mol$^{-1}$ as unit. It must be emphasised that even small changes in free energy may lead to substantial effects in binding constant values. A convenient rule of thumb is that a tenfold increase in the binding constant $K_a$ corresponds to an increase in binding affinity of only 1.4 kcal mol$^{-1}$. Note that the rate of the reaction does not depend on favourable thermodynamics, but is regulated by the kinetic rate constants described above (equation 3.2).

A thermodynamic profile of a binding reaction indicates the predominant forces that drive binding between molecules. The binding free energy value $\Delta G$ can be decomposed in changes in enthalpy ($\Delta H$) and changes in entropy ($\Delta S$):

$$\Delta G_{bind} = \Delta H - T\Delta S \tag{3.7}$$

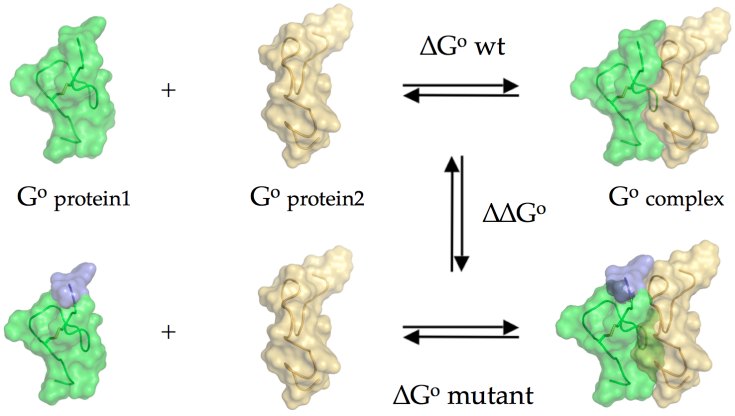


Figure 3.1: **Illustration of a binding free energy calculation of a protein-protein interaction.** $\Delta\Delta G^{\circ}$ is the relative binding free energy between the absolute binding free energy of the wild type protein complex ($\Delta G^{\circ}_{wt}$) and the absolute binding free energy where one of the proteins is mutated ($\Delta G^{\circ}_{mutant}$).

Characterisation of a binding reaction thus depends on the determination of the enthalpy and entropy components at a specified reference temperature and pressure. Every system seeks to achieve a minimum in free energy, therefore the enthalpy changes should be large and negative, while entropy changes should be large and positive. We will discuss the meaning of each thermodynamic term more in detail below.

### 3.2.3 Enthalpy

Enthalpy is defined as the thermodynamic quantity equivalent to the total heat content of a system.

$$H = U + pV \tag{3.8}$$

with $U$ the internal energy, $p$ the pressure and $V$ the volume of the system. Favourable enthalpy changes upon binding arise from stronger interactions between the binding molecules relative to interactions with solvent molecules.

$$\Delta H_{tot} = \Delta H_{coul} + \Delta H_{vdw} + \Delta H_{solv}{}^{1} \tag{3.9}$$

[1] Those terms contribute to *free energies* ($G$), where entropic effects are also taken into account.

These interactions are van der Waals interactions ($\Delta H_{vdw}$) and electrostatic interactions like hydrogen bond formations ($\Delta H_{coul}$). If the enthalpy is negative and thus favourable, then the interactions between the binding molecules are stronger than the interactions with the solvent molecules ($\Delta H_{solv}$). Otherwise, the enthalpy term will be positive and unfavourable.

We must emphasise that a quantitative amount of hydrogen bonds is not sufficient to calculate the hydrogen bond energy in $\Delta H_{coul}$. The qualitative strength of each bond depends on the optimal distance, angle and charge of the donor-acceptor pair compared to inhibitor-solvent interactions [19]. Hydrogen bonds should thus be aimed at already structured regions of the protein. Also, the formation of multiple hydrogen bonds can decrease the mobility of an inhibitor in the binding pocket, resulting in a smaller conformational entropy (see below), which indirectly improves the binding affinity between both molecules even further. Thus, the enthalpy is considered as a specific term due to directionality and proper location of hydrogen bond acceptors and donors between protein and binding partner. Finding selective drugs is therefore defined mostly by the enthalpy term and not entropy.

The solvation term, $\Delta H_{solv}$, is unfavourable for the enthalpy if interactions of the molecule with the solvent are stronger than with the binding partner. Addition of polar groups to introduce additional hydrogen bonds with the protein target does not always enhance the binding affinity of an inhibitor due to a desolvation penalty. This penalty states that positive and therefore disadvantageous enthalpies arise from transfer of polar groups from the solvent to the (usually hydrophobic) protein binding pocket. Thus, strong interactions of the molecule with the target are usually required to compensate for the unfavourable desolvation penalty of the polar groups.

### 3.2.4 Entropy

Entropy is considered as a measure of disorder of a system. The first law of thermodynamics assumes that the total energy of an isolated system is always constant. In addition, the second law of thermodynamics states that the total entropy of a system and its surroundings will always increase for a spontaneous process until a thermodynamic equilibrium is reached:

$$\frac{\Delta S}{\Delta T} \geq 0 \tag{3.10}$$

In binding reactions, entropy plays an important role for reaching the equilibrium state. A number of terms contribute to the total entropy:

$$\Delta S_{tot} = \Delta S_{solv} + \Delta S_{conf} \tag{3.11}$$

Entropy changes between biomolecules arise from two contributions, the conformational entropy ($\Delta S_{conf}$) associated with the loss of conformational degrees of freedom upon binding of both reaction partners, and the solvation entropy ($\Delta S_{solv}$) related to the burial of hydrophobic groups from the solvent.

The conformational or configurational entropy $\Delta S_{conf}$ can be explained as follows. A flexible linear ligand will adopt an ensemble of conformationally flexible molecules in solution and will therefore greatly reduce its conformational entropy upon binding. Consequently, a high number of rotational and translational degrees of freedom is unfavourable for binding. In contrast, constrained molecules with relatively few rotatable bonds will only adopt a limited amount of conformations and will consequently only have a marginal conformational entropy penalty upon binding. Hence, a possible drug optimisation strategy is the introduction of conformational restraints in the binding molecule such that the conformational space of the free and bound states is more alike. Note that residues in the protein target become restricted in their mobility as well upon binding, which is also entropically disadvantageous.

While the conformational entropy is usually unfavourable, the desolvation entropy $\Delta S_{solv}$ on the other hand is one of the most favourable contributors to binding. A number of water molecules are expelled from the protein binding site upon binding of a ligand. This desolvation process increases the degrees of freedom in the system, which consequently increases the entropy term by becoming more positive and thus more beneficial to binding. The binding molecule properties contribute to the desolvation entropy term as well because many hydrophobic forces become buried from an aqueous environment upon binding. This process, designated as the hydrophobic effect, is also an essential driving force of the folding of proteins. Here, hydrophobic residues become buried by an entropically favourable process.

In drug design, the desolvation entropy term is correlated with the hydrophobicity of the compound. To increase the potency of a potential drug molecule, it is therefore common to increase the hydrophobicity by attachment of apolar groups or replacement of polar into apolar functionalities. However, the hydrophobic character cannot increase without limit, because the molecule will become insoluble and therefore useless as a drug molecule at a certain hydrophobicity level. In

addition, some hydrophobic groups can become overexposed to the solvent even when binding to a narrow binding pocket, leading to an unfavourable solvation entropy.

Finally, we must highlight that, in contrast to enthalpy, entropy terms are non-specific due to the lack of specificity of hydrophobic forces. Hence, maximal affinity are usually only obtained when both the enthalpy and entropy changes are optimised simultaneously.

### 3.2.5 Enthalpy-entropy compensation

In contrast to the enthalpy term, the entropy component has been proven to be much easier to optimise. Enthalpy is more difficult to modulate because individual molecular interactions between drug and target like hydrogen bonds need to be known precisely. These unexpected changes are often difficult to interpret and to alleviate [20–22]. Because a number of terms are responsible to obtain a final binding affinity (as described above), entropy and enthalpy are intrinsically coupled. As a result, a notorious problem in drug optimisation strategies is that modifications to improve one term can be mitigated by an increase of other unfavourable terms. This process, called enthalpy-entropy compensation, is an infamous issue in drug optimisation studies [23–25]. Two drug molecules with different enthalpy and entropy profiles can have similar binding affinities. In fact, compounds that exhibit extremely high affinity can possess both beneficial entropy and enthalpy contributions. As a result, the total binding free energy, including the entropy, is preferably included to rank the potency of inhibitors. Except when the entropic or enthalpic nature of a series of compounds is already known in advance, studies ignoring one of these components can give misleading results, thereby disrupting a drug design optimisation trial [23, 24, 26].

## 3.3 Experimental methods

Note that absolute total free energies ($G$) cannot be measured directly and therefore only absolute differences ($\Delta G$) between states are obtained from experiments. Association constants can be measured with many experimental techniques such as fluorescence spectroscopy, ultracentrifugation, binding assays, isothermal titration calorimetry (ITC), surface plasmon resonance (SPR) and consequently the calculation of the reaction binding free energy of a reaction. In addition, the binding enthalpy can be estimated using spectroscopic techniques by performing experiments at different temperatures using the van't Hoff equation (equation 3.3).

However, these van't Hoff enthalpies are sometimes difficult to obtain and susceptible to experimental errors and artefacts compared to calorimetric enthalpies.

We will explain here shortly the principles and advantages of ITC. It is a useful technique to simultaneously obtain the stoichiometry, association constant, the enthalpy and entropy of two distinct molecules upon binding [27–29]. In ITC, a solution of dissolved molecules is injected incrementally into a reaction cell containing the binding partner. Each injection generates an amount of heat and the binding enthalpy is subsequently calculated from the amount of power required to maintain a constant temperature difference between the reaction cell and a reference cell. Interestingly, ITC is the only method to measure both enthalpy and entropy contributions coincidently, so that complexes with similar binding free energy values though different enthalpy and entropy contributions can be distinguished. In addition, ITC does not depend on immobilisation of one of the binding components, in contrast to SPR, but derives affinities from biomolecules in solutions.

SPR is an other effective method to determine binding free energies experimentally. With SPR biosensors, a molecule of interest is immobilised on a metal sensor surface. The binding partner of the immobilised molecule is subsequently injected onto the surface. The SPR instrument will measure electromagnetic wave changes due to binding of the soluble molecules with the immobilised target molecule [30, 31]. The advantage of SPR is that both kinetic and thermodynamic parameters can be extracted from a single experiment.

## 3.4 Approximate binding free energy methods

Molecular mechanics/Poisson-Boltzmann surface area (MM-PBSA) [13, 32, 33] and molecular mechanics/generalised-Born surface area (MM-GBSA) [33–35] are two similar binding free energy methods. Binding affinities can be estimated from sufficiently sampled simulation trajectories, where snapshots (i.e. conformations at specified time frames) from the molecules of interest are extracted in regular intervals. Affinities can thus be calculated from a single simulation run. That makes MM-PBSA and MM-GBSA popular methods for ranking of drugs.

These approximate methods are considered "endpoint" methods because they are restricted to conformations before and after binding, and therefore do not depend on intermediate states like exact methods. In addition, the application of a thermodynamic cycle and a continuum representation of the solvent further reduces the computational cost. Because both methods only differ in their solvent

representation, we will refer to these methods as the MM-PB/GBSA method in the discussion below.

The thermodynamic cycle and the continuum solvent models are computational "tricks" to facilitate the calculation of a binding affinity. Most energy interactions in a solvent immersed protein-ligand complex are caused by solvent-solvent interactions, which obscure the binding affinities between protein and ligand. Calculating an accurate binding free energy would require sufficiently long simulations to reach converged energy levels due to the noise caused by small variations of the solvent molecules. To this end, water and ions are stripped from the structures after the all-atom simulation, and MM-PB/GBSA applies a thermodynamic cycle to alleviate the solvent convergence issue. In this cycle, the solute is moved from a vacuum to a continuum solvent environment. By replacing an explicit water model with an implicit representation, the degrees of freedom in the solvent are reduced significantly. This simplifies the estimation of solvent binding free energies. As a result, although solvation free energy is sometimes unintentionally misspelled as "salvation" free energy (see for example Mobley *et al.* [4, 36], Tan *et al.* [37], Chen *et al.* [38], Moreira *et al.* [39] and Suenaga *et al.* [40]), MM-PB/GBSA can be regarded as a "salvation" in the field of binding free energy methods by accelerating the electrostatic calculations considerably. In aggregate, MM-PB/GBSA is several orders of magnitude faster than exact methods like FEP and TI.

### 3.4.1 Implementation

MM-PB/GBSA is one of the most well-known binding free energy methods. It estimates binding free energies by combining classical molecular mechanics, continuum solvent methods and conformational entropies. Comparable to equation 3.4, the absolute binding free energy $\Delta G_{bind}$ is derived from the free energy of the complex ($G_{AB}$) minus the receptor ($G_A$) and ligand ($G_B$) free energies. Free energy is a state function and therefore identical free energy differences can be calculated from different routes between states. This principle is applied in the thermodynamic cycle shown in Figure 3.2, where the binding free energy of each molecule is determined by a combination of enthalpic and entropic contributions.

The average free energy of a state in each environment is approximated by:

$$G_{tot} = G_{MM} + G_{solv} - TS \tag{3.12}$$

$$G_{sub-tot} = G_{MM} + G_{solv} \tag{3.13}$$

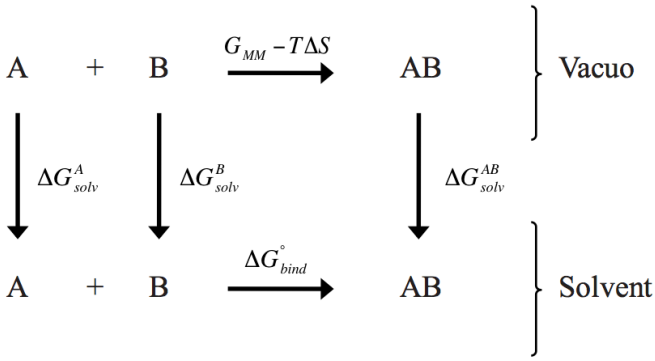


Figure 3.2: **A thermodynamic cycle applied in a typical MM-PB/GBSA setup.**

Here, the total free energy $G_{tot}$ is calculated by summation of the molecular mechanics free energy ($G_{MM}$), solvation free energy ($G_{solv}$) and configurational entropy ($-TS$). The entropy term $-TS$ is usually neglected in a standard MM-PB/GBSA approach. In that case, not a total binding free energy ($G_{tot}$) but rather a partial *subtotal* binding free energy ($G_{sub-tot}$) is estimated. Such an approximation is warranted when the conformational entropy is negligible, for example when geometrically constrained ligands are compared or the conformations of the ligands do not differ from each other significantly.

The meaning of each term in equation 3.12 is explained below.

### Molecular Mechanics

The molecular mechanics term in MM-PB/GBSA, $G_{MM}$, is simply calculated from the internal energy ($G_b$), van der Waals ($G_{vdw}$) and electrostatic interactions ($G_{coul}$) in a standard force field using no non-bonded cut-offs (see equation 2.1):

$$G_{MM} = G_b + G_{vdw} + G_{coul} \tag{3.14}$$

Because $G_{MM}$ is usually determined in the gas or vacuum phase, that term is referred to as the gas phase ($G_{gas}$) or vacuum free energy ($G_{vac}$). In principle, $G_{MM}$ can also be replaced with quantum mechanics ($G_{QM}$) if a higher precision is wanted, for example in the active site of a protein [41].

The internal bonded energy $G_b$ is ignored when complex, receptor and ligand conformations are extracted from the same simulation trajectory. We will discuss the considerations of that approach more in detail in section 3.4.3.

### Solvent models: Poisson-Boltzmann and generalised Born

The solvation free energy $G_{solv}$ is more difficult and time-consuming to calculate and is determined by transferring each system (free ligand, free receptor and complex) separately from the gas phase into the solution phase (as depicted in Figure 3.2). It is composed of polar ($G_{PB}$ or $G_{GB}$) and apolar components ($G_{SA}$):

$$G_{solv} = G_{PB/GB} + G_{SA} \tag{3.15}$$

The polar term is usually referred to as the electrostatic or polarisation energy component of the solvation free energy. We will explain the meaning of this term more in detail in this section. In principle, electrostatics can be solved using the Coulomb's law (equation 2.8). Unfortunately, this law does not consider changes in dielectrics when molecules are transferred between different environments. As a consequence, Coulomb electrostatics are unsuitable in this context. Therefore, the polar solvation energy should be calculated by a continuum method instead, such as Poisson-Boltzmann ($G_{PB}$) or generalised Born ($G_{GB}$). The basic principles of electrostatics in proteins are summarised in a number of review articles [42–47], including emphasis on implicit solvent models [48–52] like PB [53–56] and GB [57, 58].

In MM-PBSA, the polar solvation term is obtained from a numerical solution of the PB equation. This is a partial differential equation of the molecular electrostatic potential, which depends on the function and position of atomic charges, the dielectric constant in relation to the position of the charge, and the ionic strength of the medium. It yields the polarisation energy, i.e. the energy difference of transferring a molecule from the gas phase to an aqueous phase (with different dielectric constants in both phases).

The calculation can be understood by first outlining the rigorous Poisson's equation of a dielectric medium:

$$\nabla[\epsilon(\mathbf{r})\nabla\phi(\mathbf{r})] = -4\pi\rho(\mathbf{r}) \tag{3.16}$$

Here, $\epsilon$ is the space-dependent dielectric constant or relative permittivity of the medium, $\phi$ is the electrostatic potential, $\mathbf{r}$ is the position vector, $\rho(\mathbf{r})$ is the charge density in the medium and $\nabla$ is the differential operator using Cartesian coordinates.

The dielectric constant is a term that describes the polarisability of a medium, for example by applying an electric field on a system. In continuum models, the protein is considered as a continuum medium with low polarisability ($\epsilon_p = 1 - 4$, in vacuum, though even values of 20 have been reported) enclosed by the solvent with a continuum medium of high polarisability ($\epsilon_w = 78 - 80$, which is typical for water). Water has a high dielectric constant because the dipole moments of the water molecules will rapidly orient parallel to each other under the influence of the field. In contrast, a much weaker dipole moment is present in proteins despite the presence of charged and polar side chains depending on the amino acid composition. The polarisability of a protein is therefore mainly restricted to electronic redistributions, which result in a much lower dielectric constant than water. It is important to note that the protein's dielectric constant $\epsilon_p$, often termed the internal dielectric constant, is not a universal constant but is basically a parameter that depends on the specific model used [59].

If the charge distribution is written in terms of a set of fixed point charges, then solving the Poisson equation fundamentally results in the law of Coulomb (see equation 2.8). To account for the mobile ionic strength in the medium and by assuming that these ions are dispersed according to a Boltzmann distribution, a Boltzmann factor can be added to the Poisson equation, yielding the so-called nonlinear Poisson-Boltzmann formula:

$$\nabla[\epsilon(\mathbf{r})\nabla\phi(\mathbf{r})] = -4\pi\rho(\mathbf{r}) + \epsilon(\mathbf{r})\kappa^2 sinh(\phi(\mathbf{r})) \tag{3.17}$$

where $\kappa^2$ is the Debye-Hückel screening constant representing the ionic strength of the solution. The equation can be linearised to reduce the complexity and thus to facilitate the calculation:

$$\nabla[\epsilon(\mathbf{r})\nabla\phi(\mathbf{r})] = -4\pi\rho(\mathbf{r}) + \kappa^2\phi(\mathbf{r}) \tag{3.18}$$

The linearised and approximated Poisson-Boltzmann equation is sufficient for most applications and can be solved analytically by mapping a system on a three-dimensional cubic grid. The parameter values required for solving the PB equation, such as the charge density, dielectric constant and ionic strength, are assigned to each lattice point (representing biomolecule or solvent). A fine grid is necessary

to obtain accurate electrostatic values but also increases the computational cost, hence a trade-off should be found between lattice grid size and reliability of the calculation. Ultimately, the PB equation yields the electric field around a macromolecule, which is important for attraction/repulsion of ligands in many protein recognition processes. Furthermore, it can be used to estimate solvation free energies, as outlined here.

The solvation process is mimicked by transferring the molecules of interest from a dielectric continuum of low polarisability to a dielectric medium of high polarisability. Hence, the polar solvation free energy should be solved twice. Two dielectric constants yield two different electrostatic potentials; the difference between those potentials is designated by the reaction field energy:

$$\phi_{reac} = \phi_{solv} - \phi_{gas} \tag{3.19}$$

From that equation, the polar component of the solvation free energy can be written as:

$$\Delta G_{PB} = \frac{1}{2} \sum_i q_i \phi_{reac}(\mathbf{r}_i) \tag{3.20}$$

where $q_i$ is the charge assigned to each grid point $i$. This equation is implemented in the AMBER package to determine solvation free energies and in consequence binding free energies between two biomolecules. In addition, it can be calculated with stand-alone packages like DelPHI [60] and APBS [61].

A simplification of continuum treatment is the generalised Born (GB) method. GB is based on the principle that the dielectric screening of interactions between point charges is correlated with the amount of interaction of the charge with neighbouring water molecules. Due to that simplification, GB is considerably faster than PB [49], though is still able to capture as much as possible the physics of the Poisson equation. The electrostatic contributions to the solvation free energy as written in equation 3.20 can be approximated with the Born formula for a single particle with charge $q$ and radius $a$ as:

$$\Delta G_{Born} = -\frac{q^2}{2a}(1 - \frac{1}{\epsilon_{solv}}) \tag{3.21}$$

The generalised Born model can subsequently be obtained by "generalising" the Born formula over a system of many atoms of arbitrary shape. Or in other words:

multiple partial charges with respective radii are "fused" together and treated as a set of spheres:

$$\Delta G_{GB} = -\frac{1}{2}(1 - \frac{1}{\epsilon_{solv}}) \sum_{i,j}^{N} \frac{q_i q_j}{f(r_{ij}, a_{ij})} \tag{3.22}$$

with $a_{ij}$ the Born solvation radii of the atoms and $f(r_{ij}, a_{ij})$ a Gaussian function that generates the "effective Born radius" by characterising the degree of burial of the molecule. The latter term reflects the distance between the point charge and the edge of the solvent and depends not only on $a_{ij}$ but also on corresponding radii and positions of neighbouring atoms. Accurate estimation of the effective Born radius is essential to assign boundary conditions and therefore explains most differences between the developed GB models.

The GB contribution is often termed as the solvent-induced reaction field energy, because it corresponds to the favourable energy generated by the solvent in response to a charge. Due to its theoretical foundation, PB methods are considered as the benchmark and often the accuracy of GB methods is assessed by comparison with PB calculations. Most GB models show close agreement with corresponding PB calculations [62–64].[2]

### Surface Area

The apolar solvation term $G_{SA}$ is derived from the solvent-accessible surface area ($SASA$) of each molecule [65]:

$$G_{SA} = \gamma SASA + b \tag{3.23}$$

Here, $\gamma$ is an empirical surface tension constant, while $b$ is an additional empirical parameter, both derived from experimental solvation energies. In this manuscript, we applied the ICOSA technique for the $SASA$ determination [66]. The $SASA$ of complex, receptor and ligand are measured using a probe radius (usually between 1 to 1.5 Å) that determines the boundary surface between molecule and solvent. The term will fluctuate in time and can be averaged from the conformations of an equilibrated MD trajectory.

The apolar term $G_{SA}$ can be understood in terms of the hydrophobic effect, an entropically driven process. The more hydrophobic residues are buried upon

[2] We show in chapter 4 that the accuracy differences between MM-PBSA and MM-GBSA are negligible for peptide-peptide interactions.

binding (implying larger $SASA$ values), the more favourable the apolar solvation free energy. Consequently, $G_{SA}$ approximates the desolvation entropy $\Delta S_{solv}$ from equation 3.11. The proportionality of $G_{SA}$ to $SASA$ is however limited to a certain extent as it ignores subtle shape differences between molecules.

### Configurational entropy

Configurational entropy, a measure of the relative disorder in a system, requires global sampling of the conformational landscape and is therefore considered as one of the most difficult terms to estimate. Nevertheless, configurational entropy effects can contribute considerably to binding energy values [67]. There are multiple ways to approximate that entropy using simulations, albeit rather qualitatively. The method applied in this thesis is the quasi-harmonic approximation of the absolute entropy, which assumes that the fluctuations in a system can be approximated by a multivariate Gaussian probability distribution.

The true entropy of a system can be approximated from a collection of independent harmonic oscillators $i$ of corresponding frequencies $\omega_i$ using the following probability distribution:

$$S_{ho} = k_B \sum_{i}^{3n-6} \frac{\lambda}{e^{\lambda} - 1} - ln(1 - e^{-\lambda}) \tag{3.24}$$

where the eigenvalues $\lambda_i$ are obtained from corresponding quasi-harmonic frequencies $\omega_i$:

$$\lambda_i = \frac{1}{2\pi}\sqrt{\frac{k_B T}{\omega_i}} \tag{3.25}$$

These frequencies are calculated from diagonalisation of the covariance matrix ($\sigma_{ik}$) of the mass-weighted atomic coordinate fluctuations:

$$\sigma_{ik} = \sqrt{m_i m_k}\langle(x_i - \langle x_i\rangle)(x_k - \langle x_k\rangle)\rangle \tag{3.26}$$

with $m$ and $x$ respectively the masses and coordinates of the atoms in the system after least square fitting of all MD trajectory snapshots to the protein atoms.

The vibrational entropies are calculated in both bound ($AB$) and unbounded ($A$,$B$) conformations to yield the configurational entropy changes upon binding:

$$-T\Delta S = -T(S_{AB} - S_A - S_B) \tag{3.27}$$

Alternatively, a normal mode (NM) analysis can be conducted. This method derives the entropy from the second derivative of the potential energy from a single minimised conformation. A drawback of this approach is that artefacts can arise during the minimisation step, which is a prerequisite of conducting NM analysis. The approach is also restricted to a local region of the configurational space and is therefore not applicable to solvated molecules with high mobility. In contrast, quasi-harmonic analysis attempts to characterise the global configurational space of a molecular system by sampling multiple potential wells [68], yet also bears a number of disadvantages. Obtaining convergence is intrinsically complicated due to anharmonicity in the phase space. The eigenvectors of a peptide system have been shown to converge rather slowly in one microsecond of simulation time [69]. Reliable entropy estimates are more difficult to obtain for protein systems [70, 71]. In fact, it has been suggested that convergence could not be achieved for larger proteins even on a millisecond time scale [72], which implies that estimations can only provide an indication of the true entropy of larger systems.

### Summary

For clarity, we will summarise the different contributions of the MM-PB/GBSA method to the calculated binding free energy values below:

$$G_{tot} = G_{MM} + G_{solv} - TS \tag{3.28}$$

$$= G_{MM} + G_{PB/GB} + G_{SA} - TS \tag{3.29}$$

$$= G_b + G_{vdw} + G_{coul} + G_{PB/GB} + G_{SA} - TS \tag{3.30}$$

$$\text{with } G_{ele-tot} = G_{coul} + G_{PB/GB} \tag{3.31}$$

$$\text{and } G_{hyd-tot} = G_{vdw} + G_{SA} \tag{3.32}$$

Often, the total electrostatic energy of a molecule ($G_{ele-tot}$) is derived from the sum of the internal electrostatic energy ($G_{coul}$) and electrostatic solvation free energy

($G_{PB/GB}$). In addition, the total hydrophobic binding force can be obtained from the van der Waals energy ($G_{vdw}$) and apolar solvation free energy ($G_{SA}$).

### 3.4.2 Applications

A number of excellent reviews discusses the methodology and applications of MM-PB/GBSA methods in detail [13, 49, 51, 56, 73]. We will outline a selection of notable examples and applications of MM-PB/GBSA in scientific research below.

#### Overview of research applications

Due to its lower computational cost compared to "alchemical" binding free energy methods, the MM-PB/GBSA method has become increasingly popular in protein studies and drug design workflows. One of the first applications of MM-PB/GBSA was a stability study by Srinivasan *et al.* [32], who calculated the binding affinities of two helices in DNA and RNA duplexes. It was found that the B-form of DNA is clearly more favourable than the A-form, in correlation with previous experimental observations. The methodology was later extended to protein-peptide [33], protein-protein [70, 74] and protein-ligand binding interactions [75, 76].[3] In the latter approach, MM-PB/GBSA is now commonly used in conjunction with docking and MD simulations to estimate binding affinities of a series of inhibitory compounds [77, 78]. As such, the method is implemented to investigate the activity of known molecules. The method is also applicable in prospective virtual screening protocols to identify potential drug candidates. Good and weak inhibitors can be distinguished based on interaction energies with the target protein. It has been demonstrated that MM-PB/GBSA considerably outperforms current scoring functions in popular docking software packages [79]. MM-PB/GBSA can thus be implemented as a post-docking filter to allow a more rigorous and accurate ranking of a selection of docked molecules [80, 81]. MM-PB/GBSA is however significantly more demanding than basic scoring function and is therefore frequently integrated in HPC systems to accelerate the binding affinity calculations [82–84]. A recent example of an MM-PB/GBSA integrated filtering approach is the discovery of promising small-molecule anticancer drugs by Shima *et al.* [85].

While MM-PB/GBSA can be useful to seek prospective novel inhibitors, it can also be applied to tackle drug efficacy complications like the emergence of drug resistance. Wang *et al.* [86] instigated one of the first computation studies to explain

[3] In chapter 4, we apply binding free energy calculations on a peptide-peptide interaction using a modified MM-PB/GBSA setup.

the molecular basis of drug resistance of viral inhibitors against HIV-1 protease. An agreeable level of correlation with experiments was found, despite a neglect of the configurational entropy contribution. The correlation with experiments was improved in later studies by including the entropy term [87]. The lessons applied from drug resistance studies can subsequently be applied in current drug design strategies to predict resistance mutations in proteins. Safi and Lilien [88] for example have combined MM-PB/GBSA with Dead-End Elimination [89] to identify mutations that potentially diminish drug binding.

The latest advances in both hardware and software have also allowed MM-PB/GBSA investigation of much larger complexes, such as integral membrane proteins [90], transporter proteins [91] and multimeric protein complexes [92, 93]. Conformational transitions can be explored based on the computed energy levels [94], for example due to aggregation [95] or folding of proteins [96]. A novel approach is the integration of MM-PB/GBSA in protein design strategies to distinguish highly stable designer complexes from weaker ones [97, 98].

#### Energy decomposition of the binding affinity

An advantage of MM-PB/GBSA is that the energy terms can be decomposed into the most important contributions to the overall binding affinity. Hence, it can be elucidated whether binding between specific molecules is for example driven by van der Waals interactions or merely by electrostatic effects. In addition, a distinction between enthalpy and entropy components can be made, which can aid the alleviation of enthalpy-entropy compensation issues (as described in section 3.2.5). MM-PB/GBSA can thus be very useful to reveal the dominant forces and characteristics of a binding interaction.

#### Residue decomposition of the binding affinity

Another feature is that protein residues can be decomposed in terms of their contribution to the overall affinity, identifying the hot spot residues in the binding interface [66]. This is a useful application as it allows a straightforward comparison of the amino acid binding contribution with experimental site-directed mutagenesis studies. Molecular mechanics, solvation free energy and even configurational entropy contributions can be derived on a per residue level [99].[4] Note that the decomposition property can only be calculated by MM-GBSA because it is based on a fully pair-wise potential, in contrast to MM-PBSA.

[4] In chapter 6, we perform a decomposition per residue of the binding interactions, where a configurational entropy contribution is included.

#### Computational alanine scanning

A similar approach to decomposition per residue is computational alanine scanning [100, 101]. This methodology, where amino acids of the protein-protein complex are mutated to alanine, identifies hot spot residues in the binding interface [99, 102]. For example, the relative binding free energy difference of the wild type and mutated protein are calculated. The results from the approach are compared directly with experimental alanine scans, if available [103]. A disadvantage of the technique is that mutations to alanine are not always directly linked to affinity; mutations can directly destabilise the binding region, even when no close contact with the binding partner is present. A decrease in affinity might therefore be misleading if conformational rearrangements are neglected. However, alanine scanning delivers similar results as binding free energy decomposition per residue if the conformational changes are minimal [104]. Both methods are therefore usually applied in a complementary fashion to compensate their drawbacks.

### 3.4.3 Limitations, challenges and considerations

While MM-PB/GBSA has been applied accurately in a broad variety of biomolecular studies [13, 73], it must be emphasised that a number of unsuccessful results [105, 106] revealed unfortunate limitations and pitfalls of the method.

#### Length of the simulation

It has been shown that longer simulations do not necessarily result in better correlations with experimental results [64]. This might simply be a convergence issue, as in principle the most populated states have to be sampled sufficiently to obtain accurate free energy values. Converging energy values is difficult because fluctuations in both the ligand and the binding site need to be considered. The use of an inconsistent force field where errors are propagated in longer simulations would be another explanation. We must note that many initial MM-PB/GBSA studies were performed with less accurate force fields (as discussed previously in section 2.3.2), while the reliability of recent force fields have been improved to a certain extent [107]. Furthermore, protein-ligand studies depend on a subtle interplay between the ligand conformations and the rotameric states of the binding site residues. Regarding force field accuracy, particularly the derivation of ligand parameters can sometimes be questionable [108], while mixing force fields from different roots may produce inaccurate energies in the simulation system [109]. Obviously, other ligand parameters such as protonation and tautomerisation state

must be carefully assessed before relying on a MM-PB/GBSA affinity prediction [110, 111]. The configuration of the protein may not be neglected as well, as it has been demonstrated that errors in modelled protein structures can affect binding affinity predictions significantly [106].

### Simulation protocol

MM-PB/GBSA is commonly applied by extracting structures in regular intervals (designated as snapshots) from an average ensemble generated by MD or MC simulations. Although the binding free energy can in principle be calculated from just a single structure, this strategy is inherently flawed as free energies are by definition represented by all populations in the configurational space. Even small conformational perturbations can result in significantly different binding free energy values. In addition, it is not uncommon that multiple binding modes can become apparent.

An important consideration is that the complex ($AB$), receptor ($A$) and ligand ($B$) conformations should in principle be extracted from three independent simulation trajectories. Alternatively, the conformations can be derived from just one MD simulation of the complex $AB$, with individual receptor and ligand conformations basically stripped from the same trajectory. This latter approach is actually applied widely because it simplifies the calculation time considerably. In addition, binding free energies calculations reach convergence much faster than the three-trajectory approach. Internal energy contributions between complex, receptor and ligand are cancelled ($G_b$ in equation 3.14), which reduces noise in the energy levels. This assumption unfortunately implies that the conformations of the molecules may not change considerably upon binding. Consequently, using a one-simulation MM-PB/GBSA protocol is not ideal if induced fit effects or other rearrangements are expected. For example, the single trajectory approach is not recommended for estimating configurational entropies [64], which highly depend on the conformational space of the complex, receptor and ligand conformations. As such, there is a lot of debate in the literature whether the single [105] or three-simulation approach [71, 103] is preferred. Alternatively, convergence of the one-simulation approach can be increased by determination of an average binding free energy over multiple copies of the same simulation system [103, 112]. An important rule of thumb is therefore that the calculated binding free energy values can be reproduced from multiple independent simulation trajectories.[5]

[5] In chapter 4 and 6, we perform each simulation in triplicate to verify the reproducibility of the calculated affinities.

### The choice of the internal dielectric constant

Another issue of MM-PB/GBSA is the misconception between the protein's dielectric constant as a modelling parameter to determine polar solvation free energies ($\epsilon_p$) and the dielectric constant in physical microscopic ensembles ($\bar{\epsilon}$) [46, 59, 113]. While the choice of the external dielectric constant is basically determined by the solvent medium (such as water: $\epsilon_w = 80$), the choice of the internal dielectric constant is less clear due to heterogeneity of the protein structure. The internal dielectric constant parameter is in fact a misguiding term, as in reality the relative permittivity will depend on the specific region within the protein and is therefore never "constant". Hence, there is basically no universal dielectric constant that can be applied on all protein systems. In principle, the dielectric constant value should depend on the degree of molecular flexibility and the specific region within a protein. In fact, it can be argued that, depending on the amino acid composition, treating the protein as a dielectric continuum could be too deceptive. For example, a higher protein dielectric constant value can be warranted when computing protein-ligand binding free energies because surface exposed amino acid residues are fairly flexible compared to the interior of the protein. Similarly, the mobility of water molecules in protein binding pockets will be much smaller than in bulk solvent, thereby limiting the polarisability of the binding site cavity. A mixture of amino acid side chains and solvent molecules will guide the binding site electrostatics calculation and therefore a representative dielectric constant treating both protein and solvent characteristics should be applied. Consequently, MM-PB/GBSA methods may lead to different results depending on the choice of the internal dielectric constant, as shown by a protein-protein interaction study by Dong *et al.* [114]. Follow-up studies also demonstrated that the optimal choice appears to be system dependent and does not only depend on the receptor protein but also on the ligands of interest [64, 81, 115].

A number of strategies have been suggested to improve the accuracy of the protein dielectric constant. For example, charged residues are the primary driving force of the protein's dielectric constant, thus a higher dielectric constant can be found near polar residues compared to hydrophobic groups [116, 117]. Charged residues can induce a stronger electronic polarisation near the protein surface. To this end, Archontis *et al.* [118] applied a two-step pathway for a charged ligand binding event, using a low $\epsilon_p$-value for static insertion and an increased $\epsilon_p$-value for relaxation of the environment. Similarly, multiple dielectric constants have been applied to study electron transfer [119]. The principle of variable $\epsilon_p$-values has also been implemented by Moreira *et al.* [39] on an alanine scanning mutagenesis calculation. The size of the $\epsilon_p$-values was adapted to the type of the mutated amino acid, which resulted in an improved correlation with experimental alanine

scans.[6] Another notable example is the simulation of lipid membranes, where multiple $\epsilon_p$-values are justified as well. The hydrophobicity of the interior of the membrane results in a decreased polarisable relaxation compared to the solvent. Therefore, Tanizaki *et al.* [120] constructed a heterogeneous dielectric GB model, where the $\epsilon_p$-value are adapted according to the location perpendicular to the lipid bilayer. This layered dielectric constant approach was subsequently extended to integral membrane embedded proteins [121]. Finally, it can be suggested that MM-PB/GBSA calculations on amphiphilic proteins, which contain both hydrophilic and lipophilic parts, would benefit from such an apparent dielectric constant profile. An improved representation of the physical dielectric constant can thus be obtained by splitting the amphiphilic protein in multiple continuums with distinct $\epsilon_p$-values. Li *et al.* [113] have developed a smooth protein dielectric function to account for such polarisability relaxation effects.

A critical concern is that MM-PB/GBSA estimations using implicit solvent can differ significantly from explicit solvent measurements [122]. Implicit solvent might be too crude in specific cases, thus inclusion of explicit water molecules may be required [17, 123]. This is necessary when specific water molecules play an important role in the protein dynamics and function. To reduce the computational cost, the amount of solvent molecules is limited by only considering an explicit hydration shell around the ligand binding site [124] or by truncating the protein system altogether [125]. In addition, attempts have been made to include polarisability in exact binding free energy methods to improve consistency with experimental results [126].

### The importance of entropy

Small errors in the prediction of the entropy and enthalpy components can significantly impact the total binding free energy difference. The configurational entropy term is often overestimated and frequently blamed as the culprit in MM-PB/GBSA failures. Unfortunately, evaluating these entropic contributions can be cumbersome as the corresponding errors are often larger than the other binding free energy terms [127, 128]. In fact, the term is sometimes excluded voluntarily due to large standard deviation values, unrealistic entropy contributions or the excessively high computational cost depending on the system size and entropy estimation method. In contrast, neglecting entropic contributions has been confirmed to deliver less accurate relative binding free energies in a number of cases

[6] A similar strategy to improve the estimation of peptide-peptide binding free energies is applied in chapter 4.

[87, 104, 129, 130].[7] Although there are numerous attempts to improve the reliability of the entropic computations, for example with quasi-harmonic approximation [69, 131] and NM analysis adjustments [132, 133], the configurational entropy remains a difficult term to estimate.

### Comparison with other methods

MM-PB/GBSA contains a considerable amount of approximations and therefore improvement of one of the individual terms will not necessarily improve the reliability of the total binding free energy estimates. To improve the binding free energy methods, a number of studies compare different endpoint methods with each other [62–64, 106, 115, 134, 135], with other more exact approaches such as pathway [136] and exact methods [137–140], or with all three approaches altogether [141]. While many of these review articles focus merely on differences between the methods, each method has its advantages and drawbacks such as the use of approximations (in endpoint methods) or the occurrence of sampling issues (in pathway and exact methods). Hence, it must be stressed that binding free energy methods are preferrably optimised by direct correlation with experimental binding affinities.

### Conclusion

Finally, we must underline that MM-PBSA and MM-GBSA are generally used to rank the binding affinities between a series of interactions rather than give accurate predictions of the absolute binding free energies. Relative binding free energy calculations are generally more efficient due to cancellation of errors by subtracting the absolute energy values. However, even these relative differences can sometimes deviate from experimental values due to overestimation of the energy terms. This effect is usually less concerning when not accurate binding free energy values but rather distinction between good and weakly binding molecules is needed, for example in a typical drug design workflow.

[7] We demonstrate in chapter 4 that ignoring the configurational entropy contribution for the estimation of peptide-peptide free energies is not recommended.

# References


[1] Gilson, M. K. and Zhou, H.-X. (2007). Calculation of Protein-Ligand Binding Affinities. *Annu. Rev. Biophys*. 36(1):21–42.

[2] Gohlke, H. and Klebe, G. (2002). Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. *Angew. Chem. Int. Ed*. 41(15):2644–2676.

[3] Shirts, M. R., Mobley, D. L., and Chodera, J. D. (2007). Alchemical free energy calculations: Ready for prime time? *Annu. Rep. Comput. Chem*. 3:41–59.

[4] Mobley, D. L. and Dill, K. A. (2009). Binding of Small-Molecule Ligands to Proteins: "What You See" Is Not Always "What You Get". *Structure*. 17(4):489–498.

[5] Homeyer, N. and Gohlke, H. (2013). Advances in molecular dynamics simulations and free energy calculations relevant for drug design. *In Silico Drug Discov. Des*. 50–63.

[6] Sinko, W., Lindert, S., and McCammon, J. A. (2013). Accounting for receptor flexibility and enhanced sampling methods in computer-aided drug design. *Chem. Biol. Drug Des*. 81(1):41–49.

[7] Isralewitz, B., Baudry, J., Gullingsrud, J., Kosztin, D., and Schulten, K. (2001). Steered molecular dynamics investigations of protein function. *J. Mol. Graph. Model*. 19(1):13–25.

[8] Torrie, G. M. and Valleau, J. P. (1977). Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys*. 23(2):187–199.

[9] Kumar, S., Rosenberg, J. M., Bouzida, D., Swendsen, R. H., and Kollman, P. A. (1992). The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J. Comput. Chem*. 13(8):1011–1021.

[10] Roux, B. (1995). The calculation of the potential of mean force using computer simulations. *Comput. Phys. Commun*. 91(1):275–282.

[11] Leone, V., Marinelli, F., Carloni, P., and Parrinello, M. (2010). Targeting biomolecular flexibility with metadynamics. *Curr. Opin. Struct. Biol*. 20(2):148–154.

[12] Hansson, T., Marelius, J., and Aqvist, J. (1998). Ligand binding affinity prediction by linear interaction energy methods. *J. Comput. Aided Mol. Des*. 12(1):27–35.

[13] Kollman, P. A., Massova, I., Reyes, C., Kuhn, B., Huo, S., Chong, L., Lee, M., Lee, T., Duan, Y., Wang, W., Donini, O., Cieplak, P., Srinivasan, J., Case, D. A., and Cheatham, T. E. (2000). Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc. Chem. Res*. 33(12):889–897.

[14] Chodera, J. D., Mobley, D. L., Shirts, M. R., Dixon, R. W., Branson, K., and Pande, V. S. (2011). Alchemical free energy methods for drug discovery: progress and challenges. *Curr. Opin. Struct. Biol*. 21(2):150–160.

[15] Gallicchio, E. and Levy, R. M. (2011). Recent theoretical and computational advances for modeling protein-ligand binding affinities. *Adv. Protein Chem. Struct. Biol*. 85:27–80.

[16] Parenti, M. D. and Rastelli, G. (2012). Advances and applications of binding affinity prediction methods in drug discovery. *Biotechnol. Adv*. 30(1):244–250.

[17] Deng, Y. and Roux, B. (2009). Computations of Standard Binding Free Energies with Molecular Dynamics Simulations. *J. Phys. Chem. B*. 113(8):2234–2246.

[18] de Ruiter, A. and Oostenbrink, C. (2011). Free energy calculations of protein-ligand interactions. *Curr. Opin. Chem. Biol*. 15(4):547–552.

[19] Ohtaka, H. and Freire, E. (2005). Adaptive inhibitors of the HIV-1 protease. *Prog. Biophys. Mol. Biol*. 88(2):193–208.

[20] Sharp, K. (2001). Entropy-enthalpy compensation: Fact or artifact? *Protein Sci*. 10(3):661–667.

[21] Olsson, T. S. G., Ladbury, J. E., Pitt, W. R., and Williams, M. A. (2011). Extent of enthalpy-entropy compensation in protein-ligand interactions. *Protein Sci*. 20(9):1607–1618.

[22] Ferenczy, G. G. and Keserű, G. M. (2010). Thermodynamics guided lead discovery and optimization. *Drug Discov. Today*. 15(21-22):919–932.

[23] Freire, E. (2008). Do enthalpy and entropy distinguish first in class from best in class? *Drug Discov. Today*. 13(19-20):869–874.

[24] Freire, E. (2009). A Thermodynamic Approach to the Affinity Optimization of Drug Candidates. *Chem. Biol. Drug Des*. 74(5):468–472.

[25] Ladbury, J. E., Klebe, G., and Freire, E. (2009). Adding calorimetric data to decision making in lead discovery: a hot tip. *Nat. Res. Drug. Discov*. 9(1):23–27.

[26] Chodera, J. D. and Mobley, D. L. (2013). Entropy-Enthalpy Compensation: Role and Ramifications in Biomolecular Ligand Recognition and Design. *Annu. Rev. Biophys*. 42(1):121–142.

[27] Jelesarov, I. and Bosshard, H. R. (1999). Isothermal titration calorimetry and differential scanning calorimetry as complementary tools to investigate the energetics of biomolecular recognition. *J. Mol. Recogn*. 12(1):3–18.

[28] Chaires, J. B. (2008). Calorimetry and Thermodynamics in Drug Design. *Annu. Rev. Biophys*. 37(1):135–151.

[29] Torres, F. E., Recht, M. I., Coyle, J. E., Bruce, R. H., and Williams, G. (2010). Higher throughput calorimetry: opportunities, approaches and challenges. *Curr. Opin. Struct. Biol*. 20(5):598–605.

[30] Homola, J. (2003). Present and future of surface plasmon resonance biosensors. *Anal. Bioanal. Chem*. 377(3):528–539.

[31] Gopinath, S. C. B. (2010). Biosensing applications of surface plasmon resonance-based Biacore technology. *Sensor. Actuat. B-Chem*. 150(2):722–733.

[32] Srinivasan, J., Cheatham, T. E., Cieplak, P., Kollman, P. A., and Case, D. A. (1998). Continuum Solvent Studies of the Stability of DNA, RNA, and Phosphoramidate-DNA Helices. *J. Am. Chem. Soc*. 120(37):9401–9409.

[33] Massova, I. and Kollman, P. A. (2000). Combined molecular mechanical and continuum solvent approach (MM-PBSA/GBSA) to predict ligand binding. *Perspect. Drug Discovery Des*. 18(1):113–135.

[34] Still, W. C., Tempczyk, A., Hawley, R. C., and Hendrickson, T. (1990). Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc*. 112(16):6127–6129.

[35] Tsui, V. and Case, D. (2001). Theory and applications of the generalized Born solvation model in macromolecular Simulations. *Biopolymers*. 56(4):275–291.

[36] Mobley, D. L., Chodera, J. D., and Dill, K. A. (2007). Confine-and-Release Method: Obtaining Correct Binding Free Energies in the Presence of Protein Conformational Change. *J. Chem. Theory Comput*. 3(4):1231–1235.

[37] Tan, V. B. C., Zhang, B., Lim, K. M., and Tay, T. E. (2009). Explaining the inhibition of cyclin-dependent kinase 5 by peptides derived from p25 with molecular dynamics simulations and MM-PBSA. *J. Mol. Model*. 16(1):1–8.

[38] Chen, J., Zhang, S., Liu, X., and Zhang, Q. (2010). Insights into drug resistance of mutations D30N and I50V to HIV-1 protease inhibitor TMC-114: free energy calculation and molecular dynamic simulation. *J. Mol. Model*. 16(3):459–468.

[39] Moreira, I. S., Fernandes, P. A., and Ramos, M. J. (2006). Computational alanine scanning mutagenesis—An improved methodological approach. *J. Comput. Chem*. 28(3):644–654.

[40] Suenaga, A., Okimoto, N., Hirano, Y., and Fukui, K. (2012). An Efficient Computational Method for Calculating Ligand Binding Affinities. *PLoS One*. 7(8):e42846.

[41] Kaukonen, M., Söderhjelm, P., Heimdal, J., and Ryde, U. (2008). QM/MMPBSA Method To Estimate Free Energies for Reactions in Proteins. *J. Phys. Chem. B*. 112(39):12537–12548.

[42] Koehl, P. (2006). Electrostatics calculations: latest methodological advances. *Curr. Opin. Struct. Biol*. 16(2):142–151.

[43] Sharp, K. A. and Honig, B. (1990). Electrostatic Interactions in Macromolecules - Theory and Applications. *Annu. Rev. Biophys. Biophys. Chem*. 19:301–332.

[44] Dong, F., Olsen, B., and Baker, N. A. (2008). Computational methods for biomolecular electrostatics. *Methods in cell biology*. 84:843–870.

[45] Lu, B. Z., Zhou, Y. C., Holst, M. J., and McCammon, J. A. (2008). Recent progress in numerical methods for the Poisson-Boltzmann equation in biophysical applications. *Commun. Comput. Phys*. 3(5):973–1009.

[46] Warshel, A., Sharma, P. K., Kato, M., and Parson, W. W. (2006). Modeling electrostatic effects in proteins. *Biochim. Biophys. Acta*. 1764(11):1647–1676.

[47] Warshel, A. and Papazyan, A. (1998). Electrostatic effects in macromolecules: fundamental concepts and practical modeling. *Curr. Opin. Struct. Biol*. 8(2):211–217.

[48] Baker, N. A. (2005). Improving implicit solvent simulations: a Poisson-centric view. *Curr. Opin. Struct. Biol*. 15(2):137–143.

[49] Feig, M. and Brooks, C. L., III (2004). Recent advances in the development and application of implicit solvent models in biomolecule simulations. *Curr. Opin. Struct. Biol*. 14(2):217–224.

[50] Chen, J., Brooks, C. L., III, and Khandogin, J. (2008). Recent advances in implicit solvent-based methods for biomolecular simulations. *Curr. Opin. Struct. Biol*. 18(2):140–148.

[51] Wang, J., Hou, T., and Xu, X. (2006). Recent advances in free energy calculations with a combination of molecular mechanics and continuum models. *Curr. Comput. Aided Drug Des*. 2(3):287–306.

[52] Kukić, P. and Nielsen, J. E. (2010). Electrostatics in proteins and protein–ligand complexes. *Future Med. Chem*. 2(4):647–666.

[53] Fogolari, F., Brigo, A., and Molinari, H. (2002). The Poisson-Boltzmann equation for biomolecular electrostatics: a tool for structural biology. *J. Mol. Recogn*. 15(6):377–392.

[54] Simonson, T., Archontis, G., and Karplus, M. (2002). Free Energy Simulations Come of Age: Protein-Ligand Recognition. *Acc. Chem. Res*. 35(6):430–437.

[55] Simonson, T. (2001). Macromolecular electrostatics: continuum models and their growing pains. *Curr. Opin. Struct. Biol*. 11(2):243–252.

[56] Foloppe, N. and Hubbard, R. (2006). Towards predictive ligand design with free-energy based computational methods? *Curr. Med. Chem*. 13(29):3583–3608.

[57] Bashford, D. and Case, D. (2000). Generalized born models of macromolecular solvation effects. *Annu. Rev. Phys. Chem*. 51:129–152.

[58] Onufriev, A. (2008). Implicit solvent models in molecular dynamics simulations: A brief overview. *Annu. Rep. Comput. Chem*. 4:125–137.

[59] Schutz, C. N. and Warshel, A. (2001). What are the dielectric constants of proteins and how to validate electrostatic models? *Proteins*. 44(4):400–417.

[60] Honig, B. and Nicholls, A. (1995). Classical electrostatics in biology and chemistry. *Science*. 268(5214):1144–1149.

[61] Baker, N. A., Sept, D., Joseph, S., Holst, M. J., and McCammon, J. A. (2001). Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. U. S. A*. 98(18):10037–10041.

[62] Onufriev, A., Case, D. A., and Bashford, D. (2002). Effective Born radii in the generalized Born approximation: The importance of being perfect. *J. Comput. Chem*. 23(14):1297–1304.

[63] Feig, M., Onufriev, A., Lee, M. S., Im, W., Case, D. A., and Brooks, C. L. (2004). Performance comparison of generalized born and Poisson methods in the calculation of electrostatic solvation energies for protein structures. *J. Comput. Chem*. 25(2):265–284.

[64] Hou, T., Wang, J., Li, Y., and Wang, W. (2011). Assessing the Performance of the MM/PBSA and MM/GBSA Methods. 1. The Accuracy of Binding Free Energy Calculations Based on Molecular Dynamics Simulations. *J. Chem. Inf. Model*. 51(1):69–82.

[65] Sitkoff, D., Sharp, K. A., and Honig, B. (1994). Accurate Calculation of Hydration Free Energies Using Macroscopic Solvent Models. *J. Phys. Chem*. 98(7):1978–1988.

[66] Gohlke, H., Kiel, C., and Case, D. A. (2003). Insights into Protein–Protein Binding by Binding Free Energy Calculation and Free Energy Decomposition for the Ras–Raf and Ras–RalGDS Complexes. *J. Mol. Biol*. 330(4):891–913.

[67] Frederick, K. K., Marlow, M. S., Valentine, K. G., and Wand, A. J. (2007). Conformational entropy in molecular recognition by proteins. *Nature*. 448(7151):325–329.

[68] Baron, R., van Gunsteren, W. F., and Hünenberger, P. H. (2006). Estimating the configurational entropy from molecular dynamics simulations: anharmonicity and correlation corrections to the quasi-harmonic approximation. *Trends Phys. Chem*. 11:87–122.

[69] Baron, R., Hünenberger, P. H., and McCammon, J. A. (2009). Absolute Single-Molecule Entropies from Quasi-Harmonic Analysis of Microsecond Molecular Dynamics: Correction Terms and Convergence Properties. *J. Chem. Theory Comput*. 5(12):3150–3160.

[70] Gohlke, H., Kuhn, L. A., and Case, D. A. (2004). Change in protein flexibility upon complex formation: Analysis of Ras-Raf using molecular dynamics and a molecular framework approach. *Proteins*. 56(2):322–337.

[71] Basdevant, N., Weinstein, H., and Ceruso, M. (2006). Thermodynamic Basis for Promiscuity and Selectivity in Protein-Protein Interactions: PDZ Domains, a Case Study. *J. Am. Chem. Soc*. 128(39):12766–12777.

[72] Genheden, S. and Ryde, U. (2012). Will molecular dynamics simulations of proteins ever reach equilibrium? *Phys. Chem. Chem. Phys*. 14(24):8662–8677.

[73] Homeyer, N. and Gohlke, H. (2012). Free Energy Calculations by the Molecular Mechanics Poisson-Boltzmann Surface Area Method. *Mol. Inform*. 31(2):114–122.

[74] Noskov, S. Y. and Lim, C. (2001). Free energy decomposition of protein-protein interactions. *Biophys. J*. 81(2):737–750.

[75] Kuhn, B. and Kollman, P. A. (2000). Binding of a Diverse Set of Ligands to Avidin and Streptavidin: An Accurate Quantitative Prediction of Their Relative Affinities by a Combination of Molecular Mechanics and Continuum Solvent Models. *J. Med. Chem*. 43(20):3786–3791.

[76] Wang, W. and Kollman, P. A. (2001). Computational study of protein specificity: the molecular basis of HIV-1 protease drug resistance. *Proc. Natl. Acad. Sci. U. S. A*. 98(26):14937–14942.

[77] Kuhn, B., Gerber, P., Schulz-Gasch, T., and Stahl, M. (2005). Validation and Use of the MM-PBSA Approach for Drug Discovery. *J. Med. Chem*. 48(12):4040–4048.

[78] Liu, H.-Y., Grinter, S. Z., and Zou, X. (2009). Multiscale Generalized Born Modeling of Ligand Binding Energies for Virtual Database Screening. *J. Phys. Chem. B*. 113(35):11793–11799.

[79] Hou, T., Wang, J., Li, Y., and Wang, W. (2010). Assessing the performance of the molecular mechanics/Poisson Boltzmann surface area and molecular mechanics/generalized Born surface area methods. II. The accuracy of ranking poses generated from docking. *J. Comput. Chem*. 32(5):866–877.

[80] Thompson, D. C., Humblet, C., and Joseph-McCarthy, D. (2008). Investigation of MM-PBSA Rescoring of Docking Poses. *J. Chem. Inf. Model*. 48(5):1081–1091.

[81] Rastelli, G., Degliesposti, G., Del Rio, A., and Sgobba, M. (2009). Binding Estimation after Refinement, a New Automated Procedure for the Refinement and Rescoring of Docked Ligands in Virtual Screening. *Chem. Biol. Drug Des*. 73(3):283–286.

[82] Brown, S. P. and Muchmore, S. W. (2006). High-Throughput Calculation of ProteinLigand Binding Affinities: Modification and Adaptation of the MM-PBSA Protocol to Enterprise Grid Computing. *J. Chem. Inf. Model*. 46(3):999–1005.

[83] Okimoto, N., Futatsugi, N., Fuji, H., Suenaga, A., Morimoto, G., Yanai, R., Ohno, Y., Narumi, T., and Taiji, M. (2009). High-performance drug discovery: computational screening by combining docking and molecular dynamics simulations. *PLoS Comput. Biol*. 5(10):e1000528.

[84] Sadiq, S. K., Wright, D., Watson, S. J., Zasada, S. J., Stoica, I., and Coveney, P. V. (2008). Automated molecular simulation based binding affinity calculator for ligand-bound HIV-1 proteases. *J. Chem. Inf. Model*. 48(9):1909–1919.

[85] Shima, F., Yoshikawa, Y., Ye, M., Araki, M., Matsumoto, S., Liao, J., Hu, L., Sugimoto, T., Ijiri, Y., and Takeda, A. (2013). In silico discovery of small-molecule Ras inhibitors that display antitumor activity by blocking the Ras–effector interaction. *Proc. Natl. Acad. Sci. U. S. A*. 110(20):8182–8187.

[86] Wang, J., Morin, P., Wang, W., and Kollman, P. A. (2001). Use of MM-PBSA in Reproducing the Binding Free Energies to HIV-1 RT of TIBO Derivatives and Predicting the Binding Mode to HIV-1 RT of Efavirenz by Docking and MM-PBSA. *J. Am. Chem. Soc*. 123(22):5221–5230.

[87] Stoica, I., Sadiq, S. K., and Coveney, P. V. (2008). Rapid and accurate prediction of binding free energies for saquinavir-bound HIV-1 proteases. *J. Am. Chem. Soc*. 130(8):2639–2648.

[88] Safi, M. and Lilien, R. H. (2012). Efficient a Priori Identification of Drug Resistant Mutations Using Dead-End Elimination and MM-PBSA. *J. Chem. Inf. Model*. 52(6):1529–1541.

[89] Desmet, J., Maeyer, M. D., Hazes, B., and Lasters, I. (1992). The dead-end elimination theorem and its use in protein side-chain positioning. *Nature*. 356(6369):539–542.

[90] Wan, S., Flowerb, D. R., and Coveneya, P. V. (2008). Toward an atomistic understanding of the immune synapse: Large-scale molecular dynamics simulation of a membrane-embedded TCR–pMHC–CD4 complex. *Molecular Immunology*. 45:1221–1230.

[91] Collu, F., Vargiu, A. V., Dreier, J., Cascella, M., and Ruggerone, P. (2012). Recognition of Imipenem and Meropenem by the RND-Transporter MexB Studied by Computer Simulations. *J. Am. Chem. Soc*. 134(46):19146–19158.

[92] Zhao, Y., Li, W., Zeng, J., Liu, G., and Tang, Y. (2008). Insights into the interactions between HIV-1 integrase and human LEDGF/p75 by molecular dynamics simulation and free energy calculation. *Proteins*. 72(2):635–645.

[93] Tintori, C., Veljkovic, N., Veljkovic, V., and Botta, M. (2010). Computational studies of the interaction between the HIV-1 integrase tetramer and the cofactor LEDGF/p75: Insights from molecular dynamics simulations and the Informational spectrum method. *Proteins*. 78(16):3396–3408.

[94] Fogolari, F. (2005). Application of MM/PBSA colony free energy to loop decoy discrimination: Toward correlation between energy and root mean square deviation. *Protein Sci*. 14(4):889–901.

[95] Campanera, J. M. and Pouplana, R. (2010). MMPBSA Decomposition of the Binding Energy throughout a Molecular Dynamics Simulation of Amyloid-Beta (Aß1035) Aggregation. *Molecules*. 15(4):2730–2748.

[96] Combelles, C., Gracy, J., Heitz, A., Craik, D. J., and Chiche, L. (2008). Structure and folding of disulfide-rich miniproteins: Insights from molecular dynamics simulations and MM-PBSA free energy calculations. *Proteins*. 73(1):87–103.

[97] Hsieh, M.-J. and Luo, R. (2004). Physical scoring function based on AMBER force field and Poisson-Boltzmann implicit solvent for protein structure prediction. *Proteins*. 56(3):475–486.

[98] Liang, S., Li, L., Hsu, W.-L., Pilcher, M. N., Uversky, V., Zhou, Y., Dunker, A. K., and Meroueh, S. O. (2009). Exploring the Molecular Design of Protein Interaction Sites with Molecular Dynamics Simulations and Free Energy Calculations. *Biochemistry*. 48(2):399–414.

[99] Zoete, V., Meuwly, M., and Karplus, M. (2005). Study of the insulin dimerization: Binding free energy calculations and per-residue free energy decomposition. *Proteins*. 61(1):79–93.

[100] Massova, I. and Kollman, P. A. (1999). Computational Alanine Scanning To Probe Protein-Protein Interactions: A Novel Approach To Evaluate Binding Free Energies. *J. Am. Chem. Soc*. 121(36):8133–8143.

[101] Huo, S., Massova, I., and Kollman, P. A. (2002). Computational alanine scanning of the 1: 1 human growth hormone–receptor complex. *J. Comput. Chem*. 23(1):15–27.

[102] Moreira, I. S., Fernandes, P. A., and Ramos, M. J. (2006). Unravelling Hot Spots: a comprehensive computational mutagenesis study. *Theor. Chem. Acc*. 117(1):99–113.

[103] Bradshaw, R. T., Patel, B. H., Tate, E. W., Leatherbarrow, R. J., and Gould, I. R. (2010). Comparing experimental and computational alanine scanning techniques for probing a prototypical protein-protein interaction. *Protein. Eng. Des. Sel*. 24(1-2):197–207.

[104] Zoete, V. and Michielin, O. (2007). Comparison between computational alanine scanning and per-residue binding free energy decomposition for protein-protein association using MM-GBSA: Application to the TCR-p-MHC complex. *Proteins*. 67(4):1026–1047.

[105] Pearlman, D. A. (2005). Evaluating the Molecular Mechanics PoissonBoltzmann Surface Area Free Energy Method Using a Congeneric Series of Ligands to p38 MAP Kinase. *J. Med. Chem.* 48(24):7796–7807.

[106] Singh, N. and Warshel, A. (2010). Absolute binding free energy calculations: on the accuracy of computational scoring of protein-ligand interactions. *Proteins*. 78(7):1705–1723.

[107] Xu, L., Sun, H., Li, Y., Wang, J., and Hou, T. (2013). Assessing the Performance of MM/PBSA and MM/GBSA Methods. 3. The Impact of Force Fields and Ligand Charge Models. *J. Phys. Chem. B*. 117(28):8408–8421.

[108] Lemkul, J. A., Allen, W. J., and Bevan, D. R. (2010). Practical considerations for building GROMOS-compatible small-molecule topologies. *J. Chem. Inf. Model*. 50(12):2221–2235.

[109] Weis, A., Katebzadeh, K., Söderhjelm, P., Nilsson, I., and Ryde, U. (2006). Ligand Affinities Predicted with the MM/PBSA Method: Dependence on the Simulation Method and the Force Field. *J. Med. Chem.* 49(22):6596–6606.

[110] Wittayanarakul, K., Hannongbua, S., and Feig, M. (2008). Accurate prediction of protonation state as a prerequisite for reliable MM-PB(GB)SA binding free energy calculations of HIV-1 protease inhibitors. *J. Comput. Chem*. 29(5):673–685.

[111] Onufriev, A. V. and Alexov, E. (2013). Protonation and pK changes in protein–ligand binding. *Q. Rev. Biophys*. 46(02):181–209.

[112] Genheden, S. and Ryde, U. (2010). How to obtain statistically converged MM/GBSA results. *J. Comput. Chem*. 31(4):837–846.

[113] Li, M., Duan, J., Qiu, J., Yu, F., Che, X., Jiang, S., and Li, L. (2013). 3-Hydroxyphthalic Anhydride-Modified Human Serum Albumin as a Microbicide Candidate Against HIV Type 1 Entry by Targeting Both Viral Envelope Glycoprotein gp120 and Cellular Receptor CD4. *AIDS Res. Hum. Retroviruses*. 29(11):1455–1464.

[114] Dong, F., Vijayakumar, M., and Zhou, H.-X. (2003). Comparison of Calculation and Experiment Implicates Significant Electrostatic Contributions to the Binding Stability of Barnase and Barstar. *Biophys. J*. 85(1):49–60.

[115] Genheden, S. and Ryde, U. (2012). Comparison of end-point continuum-solvation methods for the calculation of protein-ligand binding free energies. *Proteins*. 80(5):1326–1342.

[116] Simonson, T. and Perahia, D. (1995). Internal and interfacial dielectric properties of cytochrome c from molecular dynamics in aqueous solution. *Proc. Natl. Acad. Sci. U. S. A*. 92(4):1082–1086.

[117] Pitera, J. W., Falta, M., and van Gunsteren, W. F. (2001). Dielectric Properties of Proteins from Simulation: The Effects of Solvent, Ligands, pH, and Temperature. *Biophys. J*. 80(6):2546–2555.

[118] Archontis, G. and Simonson, T. (2001). Dielectric Relaxation in an Enzyme Active Site: Molecular Dynamics Simulations Interpreted with a Macroscopic Continuum Model. *J. Am. Chem. Soc*. 123(44):11047–11056.

[119] Rocchia, W., Alexov, E., and Honig, B. (2001). Extending the Applicability of the Nonlinear PoissonBoltzmann Equation: Multiple Dielectric Constants and Multivalent Ions. *J. Phys. Chem. B*. 105(28):6507–6514.

[120] Tanizaki, S. and Feig, M. (2005). A generalized Born formalism for heterogeneous dielectric environments: Application to the implicit modeling of biological membranes. *J. Chem. Phys*. 122(12):124706.

[121] Tanizaki, S. and Feig, M. (2006). Molecular Dynamics Simulations of Large Integral Membrane Proteins with an Implicit Membrane Model. *J. Phys. Chem. B*. 110(1):548–556.

[122] Godschalk, F., Genheden, S., Söderhjelm, P., and Ryde, U. (2013). Comparison of MM/GBSA calculations based on explicit and implicit solvent simulations. *Phys. Chem. Chem. Phys*. 15(20):7731–7739.

[123] Wong, S., Amaro, R. E., and McCammon, J. A. (2009). MM-PBSA Captures Key Role of Intercalating Water Molecules at a ProteinProtein Interface. *J. Chem. Theory Comput*. 5(2):422–429.

[124] Maffucci, I. and Contini, A. (2013). Explicit Ligand Hydration Shells Improve the Correlation between MM-PB/GBSA Binding Energies and Experimental Activities. *J. Chem. Theory Comput*. 9(6):2706–2717.

[125] Genheden, S. and Ryde, U. (2012). Improving the Efficiency of Protein–Ligand Binding Free-Energy Calculations by System Truncation. *J. Chem. Theory Comput*. 8(4):1449–1458.

[126] Jiao, D., Golubkov, P. A., Darden, T. A., and Ren, P. (2008). Calculation of protein–ligand binding free energy by using a polarizable potential. *Proc. Natl. Acad. Sci. U. S. A*. 105(17):6290–6295.

[127] Carlsson, J. and Aqvist, J. (2005). Absolute and Relative Entropies from Computer Simulation with Applications to Ligand Binding. *J. Phys. Chem. B*. 109(13):6448–6456.

[128] Singh, N. and Warshel, A. (2010). A comprehensive examination of the contributions to the binding entropy of protein-ligand complexes. *Proteins*. 78(7):1724–1735.

[129] Xu, Y. and Wang, R. (2006). A computational analysis of the binding affinities of FKBP12 inhibitors using the MM-PB/SA method. *Proteins*. 64(4):1058–1068.

[130] Kar, P., Gopal, S. M., Cheng, Y.-M., Predeus, A., and Feig, M. (2013). PRIMO: A Transferable Coarse-Grained Force Field for Proteins. *J. Chem. Theory Comput*. 9(8):3769–3788.

[131] Numata, J., Wan, M., and Knapp, E.-W. (2007). Conformational entropy of biomolecules: beyond the quasi-harmonic approximation. *Genome informatics. International Conference on Genome Informatics*. 18:192–205.

[132] Kongsted, J. and Ryde, U. (2008). An improved method to predict the entropy term with the MM/PBSA approach. *J. Comput. Aided Mol. Des*. 23(2):63–71.

[133] Genheden, S., Kuhn, O., Mikulskis, P., Hoffmann, D., and Ryde, U. (2012). The Normal-Mode Entropy in the MM/GBSA Method: Effect of System Truncation, Buffer Region, and Dielectric Constant. *J. Chem. Inf. Model*. 52(8):2079–2088.

[134] Srivastava, H. K. and Sastry, G. N. (2012). Molecular Dynamics Investigation on a Series of HIV Protease Inhibitors: Assessing the Performance of MM-PBSA and MM-GBSA Approaches. *J. Chem. Inf. Model*. 52(11):3088–3098.

[135] Genheden, S. and Ryde, U. (2011). Comparison of the Efficiency of the LIE and MM/GBSA Methods to Calculate Ligand-Binding Energies. *J. Chem. Theory Comput*. 7(11):3768–3778.

[136] Lee, M. S. and Olson, M. A. (2006). Calculation of Absolute Protein-Ligand Binding Affinity Using Path and Endpoint Approaches. *Biophys. J*. 90(3):864–877.

[137] Laitinen, T., Kankare, J. A., and Peräkylä, M. (2004). Free energy simulations and MM-PBSA analyses on the affinity and specificity of steroid binding to antiestradiol antibody. *Proteins*. 55(1):34–43.

[138] Martins, S. A., Perez, M. A. S., Moreira, I. S., Sousa, S. F., Ramos, M. J., and Fernandes, P. A. (2013). Computational Alanine Scanning Mutagenesis: MM-PBSA vs TI. *J. Chem. Theory Comput*. 9(3):1311–1319.

[139] Guimarães, C. R. W. and Mathiowetz, A. M. (2010). Addressing Limitations with the MM-GB/SA Scoring Procedure using the WaterMap Method and Free Energy Perturbation Calculations. *J. Chem. Inf. Model*. 50(4):547–559.

[140] Guimarães, C. R. W. (2011). A Direct Comparison of the MM-GB/SA Scoring Procedure and Free-Energy Perturbation Calculations Using Carbonic Anhydrase as a Test Case: Strengths and Pitfalls of Each Approach. *J. Chem. Theory Comput*. 7(7):2296–2306.

[141] Charlier, L., Nespoulous, C., Fiorucci, S., Antonczak, S., and Golebiowski, J. (2007). Binding free energy prediction in strongly hydrophobic biomolecular systems. *Phys. Chem. Chem. Phys*. 9(43):5761–5771.

# Part III

# Results

**Chapter 4 | Optimisation of VIRIP**

In this chapter, we suggest and successfully obtain a number of improved VIRIP derivatives by implementing a peptide-peptide optimised MM-PBSA virtual screening approach. The development and efficacy of the improved method is explained. In addition, the interaction of VIRIP with its target, gp41 FP, is discussed in detail and we outline a number of future directions.

The contents of this chapter have been published in:

Venken, T., Krnavek, D., Münch, J., Kirchhoff, F., Henklein, P., De Maeyer, M., and Voet, A. (2011). An optimized MM/PBSA virtual screening approach applied to an HIV-1 gp41 fusion peptide inhibitor. *Proteins*. 79(11):3221–3235

**Chapter 5 | The flexibility of HIV-1 FP in solution**

Although the results from the previous chapter are promising, the lack of structural information of the target hinders additional optimisations. Therefore, we applied multiscale MD simulations to explore the conformational ensemble of the gp41 FP in solution. We discuss the measured secondary structure properties in detail and implement a secondary structure clustering method. A number of future prospects is outlined in detail.

The contents of this chapter have been published in:

Venken, T., Voet, A., De Maeyer, M., De Fabritiis, G., and Sadiq, S. K. (2013). Rapid Conformational Fluctuations of Disordered HIV-1 Fusion Peptide in Solution. *J. Chem. Theory Comput*. 9(7):2870–2874

## Chapter 6 | Scrutiny of the Rev multimerisation

In the last result chapter, we study the multimerisation of the HIV-1 Rev protein using a computational methodology. The binding affinities between individual Rev monomers are explained in detail and a number of hot spot residues are identified and compared to experimental results. The results of this chapter form the basis for virtual screening of specific Rev multimerisation inhibitors.

The contents of this chapter have been published in:

Venken, T., Daelemans, D., De Maeyer, M., and Voet, A. (2012). Computational investigation of the HIV-1 Rev multimerization using molecular dynamics simulations and binding free energy calculations. *Proteins*. 80(6):1633–1646

# Chapter 4

# Optimisation of VIRIP

"I Want To Decompose"

Whistling - The Hickey Underworld

---

This chapter is an adapted reprint of the article:

Venken, T., Krnavek, D., Münch, J., Kirchhoff, F., Henklein, P., De Maeyer, M., and Voet, A. (2011). An optimized MM/PBSA virtual screening approach applied to an HIV-1 gp41 fusion peptide inhibitor. *Proteins*. 79(11):3221–3235.

The original introduction and materials and methods section have been shortened to avoid repetition with previous thesis chapters. I performed all the simulations, binding free energy calculations and most of the analyses. Suggested peptides were tested in the lab of prof. dr. F. Kirchhoff and prof. dr. J. Münch (Materials and methods can be found in the full paper). I wrote the paper with adjustments and additions of the co-authors.

---

## 4.1 Summary

VIRus Inhibitory Peptide (VIRIP), a 20 amino acid peptide, binds to the FP of HIV-1 gp41 and blocks viral entry. VIRIP derivatives with improved antiviral activity have been developed, and one of those derivatives has recently proven effective and safe in a phase I/II clinical trial. Here, MD simulations were executed in combination with MM-PBSA free energy calculations to explore the binding interaction between VIRIP derivatives and gp41 FP. A promising correlation between antiviral activity and simulated binding free energy was established thanks to restriction of the flexibility of the peptides, inclusion of configurational entropy calculations, and the use of multiple internal dielectric constants for the MM-PBSA calculations depending on the amino acid sequence. Based on these results, a virtual screening experiment was carried out to design VIRIP variants with further improved antiretroviral activity. A selection of peptides was tested for inhibitory activity and several VIRIP derivatives were identified with significantly enhanced

activity compared to the reference peptides. The results demonstrate that computational modelling strategies using an adapted MM-PBSA methodology improve the accuracy of binding free energy calculations of peptide complexes compared to the classic MM-PBSA protocol. As such, this virtual screening approach generated HIV-1 gp41 FP inhibitors with improved antiviral activity that could be useful for future clinical applications.

## 4.2 Introduction

As outlined previously in section 1.4.2, VIRIP and its derivatives offer an alternative antiretroviral strategy by inhibiting the anchoring event of the gp41 FP. The published NMR structure of the optimised VIR-165 derivative in complex with the FP [1] provides a paradigm for *in silico* optimisation. Here, we report an improved methodology using MD simulations and the MM-PBSA approach to investigate peptidic interactions. The method consists of the implementation of restraints, inclusion of configurational entropy calculations, and the use of multiple internal dielectric constants depending on the sequence of the peptide. To our knowledge, this is the first time that a peptide-peptide interaction has been analysed using the MM-PBSA method. The improved procedure forms the basis for the design of enhanced gp41 FP inhibitors to interfere with the HIV-1 entry process. In fact, a significant correlation between antiviral activity and simulated binding free energy was found, allowing the design of VIRIP derivatives with improved antiretroviral activity based on virtual screening.

## 4.3 Materials and methods

### 4.3.1 System preparation

Coordinates for all VIRIP derivatives were based on the NMR structure of VIR-165 (Protein Databank code: 2JNR) [1]. Amino acid point mutations were created using the Brugel package [2] by retaining the dihedral angles of the side chain followed by 1000 steps of conjugated gradient energy minimisation using a CHARMM-based force field [3].

MD simulations were performed with the GROMACS package, version 4.0.7 [4] using the OPLS/AA force field [5]. Each complex was placed in a trigonal box and filled with TIP3P water molecules [6]. Distances between the complex and the edge of the box were at least 0.85 nm. When required, chloride or sodium counterions

were added to neutralise the charges of the complex. Following conjugate gradient energy minimisation, each complex was equilibrated during 100 ps at constant pressure (NPT ensemble). During this equilibration phase, both peptides were fixed with harmonic position restraints on all heavy protein atoms with a force of 1000 kJ $mol^{-1}$ $nm^{-2}$. The water molecules were allowed to adapt to the complex, thus filling possible cavities and eliminating clashes. Next, three separate full production MD simulations of 10 ns each were executed on all complexes. In these simulations, three different conditions were applied: no restraints, "weak" restraints (a combination of dihedral restraints of the backbone atoms and position restraints only on the $C_\alpha$ atoms with a force of 100 kJ $mol^{-1}$ $nm^{-2}$), and "strong" restraints (position restraints only on the $C_\alpha$ atoms with a force of 1000 kJ $mol^{-1}$ $nm^{-2}$). For comparative reasons, an equilibration and production MD simulation was performed on the VIR-165:FP complex using NOE distance restraints similar to the MD setup for the original NMR determination [1].

During the simulations, periodic boundary conditions and the Berendsen thermostat at 298 K and barostat at 1 bar were applied [7]. Temperature and pressure were maintained with coupling constants $\tau_T$ = 0.1 ps and $\tau_p$ = 1 ps. Atom bonds were constrained using the LINCS algorithm [8] with a 2 ps integration step. Electrostatic interactions were calculated using the particle mesh Ewald summation method [9]. For the short-range van der Waals interactions, a cut-off distance of 1.4 nm was applied. Each system contained roughly 8500 atoms.

### 4.3.2 Binding free energy calculations

An average ensemble of structures was generated with MD simulations in explicit solvent as described above. Five hundred snapshots were extracted during the last 5 ns of the production MD trajectory (i.e., at 10 ps intervals), while removing solvent and counterions. Next, MM-PBSA calculations were performed using the AMBER 8 package [10] using this ensemble of structures. The MM-PBSA approach was used to determine the binding free energy of VIRIP and FP, while MM-GBSA using the Onufriev model [11] was used for binding free energy decomposition of the binding interaction. The implementation as reviewed in section 3.4.1 was used for both methods. Regarding the configuration entropy calculations, the structures used to construct the covariance fluctuation matrices of receptor, ligand, and peptidic complex were all extracted from the same MD trajectory of the peptide complex, identical to the MM-PBSA calculations. As such, rotational or translational changes upon binding are not taken into account. Thus, entropy calculations are restricted to the dynamics between FP and VIRIP in the bound state, where large entropic costs indicate that the dynamics of both peptides are

highly associated. The entropy estimations were calculated during the whole 10 ns of the production MD run using all frames. As entropy is a time-dependent parameter, 10 ns were chosen to obtain as much convergence as possible.

### 4.3.3 Methodology of the binding free energy calculations

The MM-PBSA method depends on the conformation of the studied molecules. Since structural changes have a large effect on the overall energetic profile, too much flexibility will increase the uncertainty of the binding free energy values and configurational entropy predictions. The high flexibility of the VIR-165:FP complex was confirmed from initial MD simulations (see Figure 4.1). To overcome this inherent flexibility, it is possible to apply restraints to prevent a strong deviation from the starting structure. NOE distance restraints were used for the initial determination of the NMR-structure of the VIR-165:FP complex [1]. The implementation of these NOE distance restraints was utilised for the VIR-165:FP complex, but is not applicable for the mutated VIRIP derivatives. Therefore, it was investigated whether alternative methods can be applied to simulate all VIRIP derivatives. As such, other restraining methods were explored: (i) position restraints, (ii) dihedral restraints, and (iii) a combination of position and dihedral restraints. Different restraining strengths were tested: (i) "strong" restraints of 1000 kJ $mol^{-1}$ $nm^{-2}$ on each $C_\alpha$ atom and (ii) "weak" restraints of 100 kJ $mol^{-1}$ $nm^{-2}$ on each $C_\alpha$ atom. A gradual decrease of the backbone RMSD profile is observed when increasing the restraining force (Figure 4.1A).

In the end, a combination of dihedral restraints on the backbone atoms and position restraints on the $C_\alpha$ atoms proved to be ideal. For these restraints, position restraint forces of 100 kJ $mol^{-1}$ $nm^{-2}$ on each $C_\alpha$ atom and dihedral restraint forces of 100 kJ $mol^{-1}$ $rad^{-2}$ on $\phi$ and $\psi$ angles of the backbone were applied, thereby finding a balance between flexibility and extreme deviation from the starting structure. Of note, side chain atoms remained unrestrained during all production simulation steps, allowing interfacial adaptation between both peptides and the calculation of the side chain configurational entropy contribution. As a remark, the backbone RMSD of the VIR-165:FP complex using the "weak" restraints agrees with the RMSD of the same complex using NOE distance restraints (Figure 4.1B). Thus, combination of "weak" dihedral and position restraints reproduces the original NMR structure, allowing the usage of these restraint settings. A representation of the effect of the different restraints conditions on VIR-165 and gp41 FP is shown in Figure 4.3A.

A second adaptation in the methodology is the inclusion of configurational entropy calculations, which are often neglected due to efficiency reasons and possible inac-

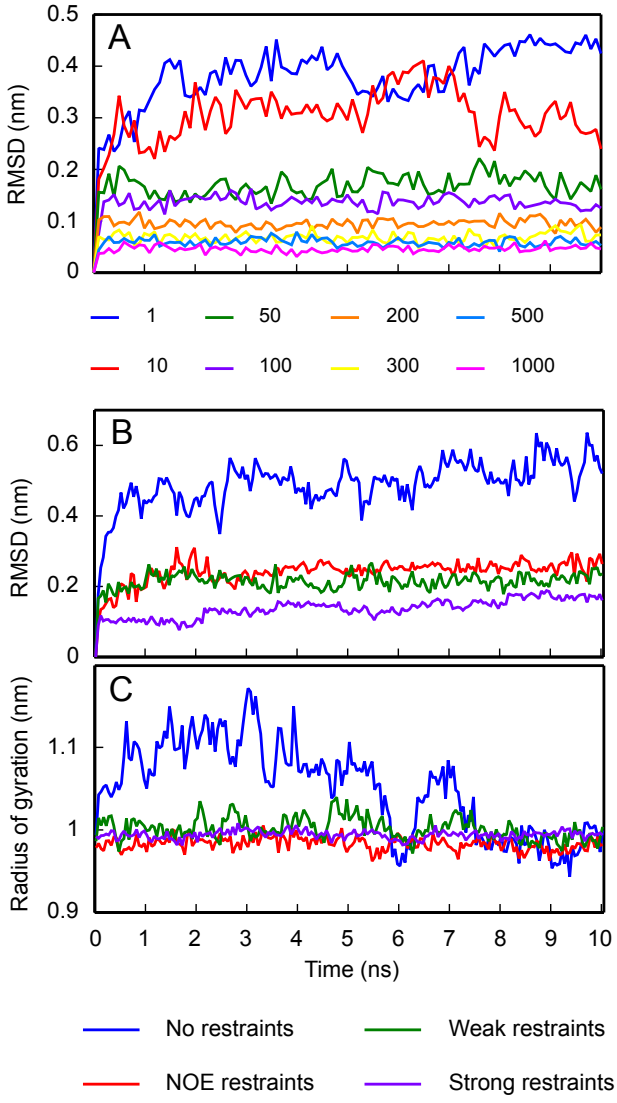


Figure 4.1: **Structural analysis of the MD simulations.** (A) Root mean square deviation (RMSD) profile of the VIR-165:FP complex when increasing position restraint strength (given in kJ $^{-1}$ nm$^{-2}$) on $C_\alpha$ atoms and dihedral restraint strength (given in kJ mol$^{-1}$ rad$^{-2}$) on $\phi$ and $\psi$ angles of the backbone. (B) RMSD of the backbone atoms of VIR-165 and (C) radius of gyration of the VIR-165:FP complex during 10-ns MD simulations. Results for four different conditions (unrestrained, NOE distance restraints, weak restraints, and strong restraints) are shown.

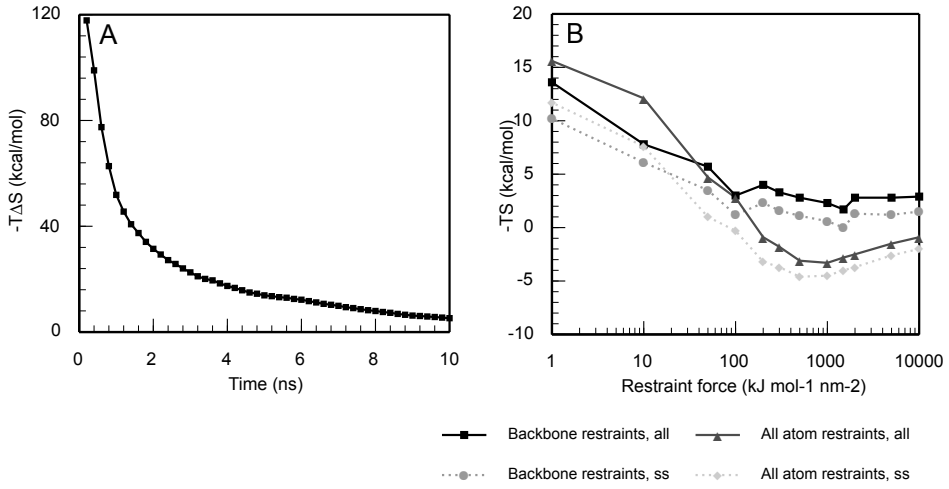


Figure 4.2: **Examination of configurational entropy estimates.** (A) Example of the convergence of the entropy part of the binding free energy of VIR-165 with gp41 FP. The configurational entropy was estimated with a quasi-harmonic approximation during a 10-ns MD simulation. (B) Effect of configurational entropy difference between complex, FP, and VIR-165 using restraint strength on all atoms or on the backbone atoms. Entropy estimates were performed on all atoms (designated as "all") or on only the side chain atoms (designated as "ss").

curacies [12–14]. However, the flexible nature of the VIR-165:FP peptide complex requests the inclusion of configurational entropy. An initial 1-ns MD simulation resulted in entropy estimates over 50 kcal mol$^{-1}$, but longer simulations (>10 ns) of the VIR-165:FP complex resulted in convergence of the entropical contribution (Figure 4.2A). This has already been observed in simulations on the PDZ domains [12], demonstrating the necessity of long simulations to obtain convergence of the entropic contribution using the quasi-harmonic analysis method. Although entropy estimates did not converge entirely after 10 ns, it is clear from Figure 4.2A that the $-T\Delta S$ value is almost converged, and thereby can provide an estimate of the binding entropy difference. Furthermore, this research is focused mainly on a relative comparison of derivatives and less on an absolute reproduction of entropic contributions.

The restraints are implemented only on the backbone atoms of the peptides, while side chain atoms are unrestrained and thus able to explore the conformational space extensively. As such, the entropy estimations in the restraint conditions are mainly a representation of the amino acid side chain motions, while the backbone only contributes partially due to the restricted mobility. Nevertheless, a major consequence of the restraint implementation is the direct effect on the flexibility of the system. Applying restraints influences the configurational entropy predictions

one wishes to evaluate. Therefore, additional calculations were performed to verify the effects of the restraint settings on the entropy estimations. In Figure 4.2B, the effect of different restraint settings on the configurational entropy estimations was explored. Two types of restraints settings were used, restraining all atoms in the system or restraining only the $C_\alpha$ atoms and dihedral backbone angles. Entropy calculations were performed on all atoms or only the side chain atoms. Restraint forces range from very weak (1 kJ $^{-1}$ nm$^{-2}$, which corresponds to an almost unrestrained condition) to very strong (10,000 kJ $^{-1}$ nm$^{-2}$, corresponding to only minimal flexibility in the peptide complex). As is visible, the restraint settings show that the side chain entropy difference between VIR-165:FP and its constituents correlates closely with the entire protein entropy difference, indicating that the backbone barely contributes, as was expected. In addition, increasing the restraint force on all atoms results in an exponential decrease of the entropy difference. This difference increases after reaching a restraint force of 1000 kJ $^{-1}$ nm$^{-2}$. In contrast, restraining only the backbone conformation results in a much weaker dependency on the restraint settings. In fact, the entropy difference no longer reduces significantly upon increasing the restraint force beyond 100 kJ $^{-1}$ nm$^{-2}$. This shows that exclusively restraining the backbone atoms results only in a limited loss of thermodynamic information compared to restraining the entire protein conformation. However, some information is lost, so the configurational entropy contribution calculated here should be considered as a virtual screening value instead as an absolute thermodynamic value.

A third adaptation is the use of multiple dielectric constants for the protein dependent on the sequence of each VIRIP derivative. As described in chapter 3, the dielectric constant of a protein is not a universal parameter, and commonly different values are used ranging from 1 to 4 (and even higher) depending on the methodology or models used [15–17]. It has been shown before that the use of different dielectric constants can drastically enhance the reliability of $\Delta\Delta G$ values using MM-PBSA. This approach, however, has solely been applied in alanine scanning simulations, and a procedure based on the same principle for other mutations must consequently be implemented. Different degrees of relaxation are present in a protein upon mutation of an amino acid to an alanine. Ramos and coworkers used the following internal dielectric constants: two for apolar residues (Val, Leu, Ile, Phe, Met, and Trp), three for polar residues (Asn, Gln, Cys, Tyr, Ser, and Thr) and four for charged residues (Asp, Glu, Lys, Arg, and His) [17]. They obtained an excellent agreement between their calculated relative binding free energy values and the experimental results. In the present study, multiple amino acids are mutated and an average dielectric constant was calculated when multiple mutations were present. Ala, Pro, and Gly mutations were added to the category with value $\epsilon_p = 2$. When no alanine substitution was present, the dielectric differ-

ence between two amino acids based on their dielectric categories was subtracted by the following rule: $\epsilon_{newvalue} = \epsilon_{oldvalue} - (\epsilon_{originalresidue} - \epsilon_{mutatedresidue})$. As such, mutation of a residue from a lower category results in an increase of the internal dielectric constant, while mutation of a residue from a higher category decreases the internal dielectric constant, analogous to the alanine scanning methodology [17]. This procedure has the advantage that multiple amino acid substitutions can be investigated and that the analysis should not be restricted to alanine scans. This implementation is justified in our case because of the small size of the VIRIP derivatives, since small amino acid sequence changes can have a substantial effect on the electric environment at the interface, and thus on the accuracy of the binding free energy estimations.

### 4.3.4 Virtual screening

Conservative mutations of specific residues were proposed to construct additional VIRIP derivatives, for example, replacement of hydrophobic residues like Val by a larger Leu residue. Some mutations were based on biochemical intuition after structural investigation of the NMR structure of VIR-165, for example, VIR-165_F12L_A13F. In addition, information from the computational alanine scanning and binding free energy decomposition (as explained above) was used to suggest mutations. Certain positions were not altered, for example, the cysteine bridge in VIR-165 and most prolines, as these mutations could have a drastic effect on the conformation of the peptide. Polar and charged residues were left unchanged (e.g. Lys16); or changed to a homologue residue. Initially, only single mutations were proposed. In total, 81 *in silico* single mutations were tested: 35 based on VIR-165 and 46 based on VIR-175. From these results, 40 additional double and triple mutants were suggested based on VIR-165. All mutations were constructed with the Brugel package [2] and simulations were carried out as described above. For reasons of accuracy, all calculations were run in triplicate using a random seed for each MD-simulation.

## 4.4 Results and discussion

### 4.4.1 Initial analysis of the VIR-165:FP peptide complex structure

Estimating the binding free energy and thereby accurately distinguishing strong from weak-binding flexible peptides is not as straightforward as for large pro-

tein–ligand and protein–protein complexes. Extensive secondary structure elements in large proteins help to stabilise the overall structure. In contrast, the structural prediction of the conformation of small peptides is significantly more difficult, since a peptide can adopt a whole spectrum of conformations. Thus, there is a need to adapt the simulation settings in order to develop a more reliable predictive model. Such a model should calculate the binding free energy of each VIRIP derivative with the gp41 FP with significant precision. To verify the flexibility of the peptide complex, a structural analysis of the VIR-165:FP interaction was performed. Based on the visual inspection of the NMR structure of VIR-165 in complex with the gp41 FP, it can be deduced that both short peptides have no clear secondary structure. Only a disulphide bridge in certain optimised VIRIP derivatives introduces some rigidity. This assumption was confirmed by a secondary structure prediction program [18], showing that VIR-165 has an average coil probability of 80%. The gp41 FP has more secondary structure, with an average a-helix probability of 40%, mostly spanning residues Gly5 to Leu12, while the average coil probability is 53%. In addition, initial MD simulations showed high flexibility of the peptide complex in unrestrained conditions, although no drastic unfolding occurs. This was verified by calculating the RMSD of the backbone atoms of VIR-165 (Figure 4.1B) and the radius of gyration of the peptide complex (Figure 4.1C). Interestingly, the NMR determination of the VIR-165:FP structure relied on NOE distance restraints, and in principle these could have been applied in our simulations. However, these restraints are only applicable in the VIR-165:FP complex and not in derivatives. Therefore, we explored different settings of dihedral restraints and position restraints of the backbone atoms, while keeping the side chain atoms unrestrained.

### 4.4.2 Restrained MM-PBSA simulations including entropy calculations

A set of VIRIP derivatives comprising only natural amino acids was selected from a previous study [1] to validate the MM-PBSA setup as explained above. Most tested VIRIP derivatives contain a disulphide bridge between Cys6-Cys11, as in, for example, VIR-165 [1]. Derivatives without a disulphide bridge were also selected. However, derivatives with disulphide bridges at other positions were not included since NMR structures of these peptides are not available, and hence a precise conformation is not known. Using these criteria, 29 derivatives were selected for further analysis (see Table 4.1). For each VIRIP derivative, restrained simulations were conducted in order to reduce the flexibility during the MD simulations with inclusion of configurational entropy calculations. The simulated

binding free energy values were compared to experimental $IC_{50}$ values, which were converted to relative free energy values using the following relationship [19]:

$$\Delta\Delta G = RTln(IC_{50}mut/IC_{50}wt) \tag{4.1}$$

where R is the ideal gas constant and T is the temperature in K. The antiviral activity as measured by the $IC_{50}$ of the tested derivatives ranges from 0.18 $\mu$M to inactive ($\gg$ 100 $\mu$M) [1].

Three different restraining strengths were tested for all 29 VIRIP derivatives (unrestrained, weakly, and strongly restrained). The effect of these restraint conditions on the RMSD of the backbone atoms of VIR-165 is shown in Figure 4.1B, while a visual representation of the VIR-165:FP complex flexibility is presented in Figure 4.3A. As can be seen in Figure 4.3B, low correlation was found between the experimental and theoretical values when unrestrained conditions were applied. Apparently, this condition results in too many degrees of freedom to allow a reliable determination of the binding free energy. In addition, configurational entropy calculations did not converge due to relatively high peptide flexibility. Although configurational entropy predictions were possible when applying strong restraints, no correlation was found. These settings result in rigidity preventing strong adaptation to the backbone conformation of the VIRIP derivatives. In contrast, a promising correlation between experimental and theoretical values is found in the case of weak restraints ($R^2$ = 0.40). These simulation settings allow a balance between rigidity and flexibility of the VIRIP derivatives in complex with FP.

VIRIP acts exterior of the cell and therefore factors such as cellular permeability, stability, and other interferences of the cellular environment do not influence the activity [1]. As such, a correlation between the simulated free energy of binding values and *in cellulo* antiviral activity is expected and has been found in the experimental data. Although the implementation of restraints during the MD simulations was obligatory, this procedure might skew the results for some derivatives, explaining the good but imperfect correlation. Furthermore, the force field, limitations of the entropy calculations, and other approximations inherent to the MM-PBSA methodology (such as the implicit solvent model) cause deviations. As such, to evaluate the predictive power of the MM-PBSA method, receiver operator characteristic (ROC) curves were plotted [20]. A ROC curve corresponds to the enrichment of the method and forms an association between sensitivity (defined by the true positive rate) and specificity (defined by the false positive rate). ROC analysis revealed a promising predictivity for

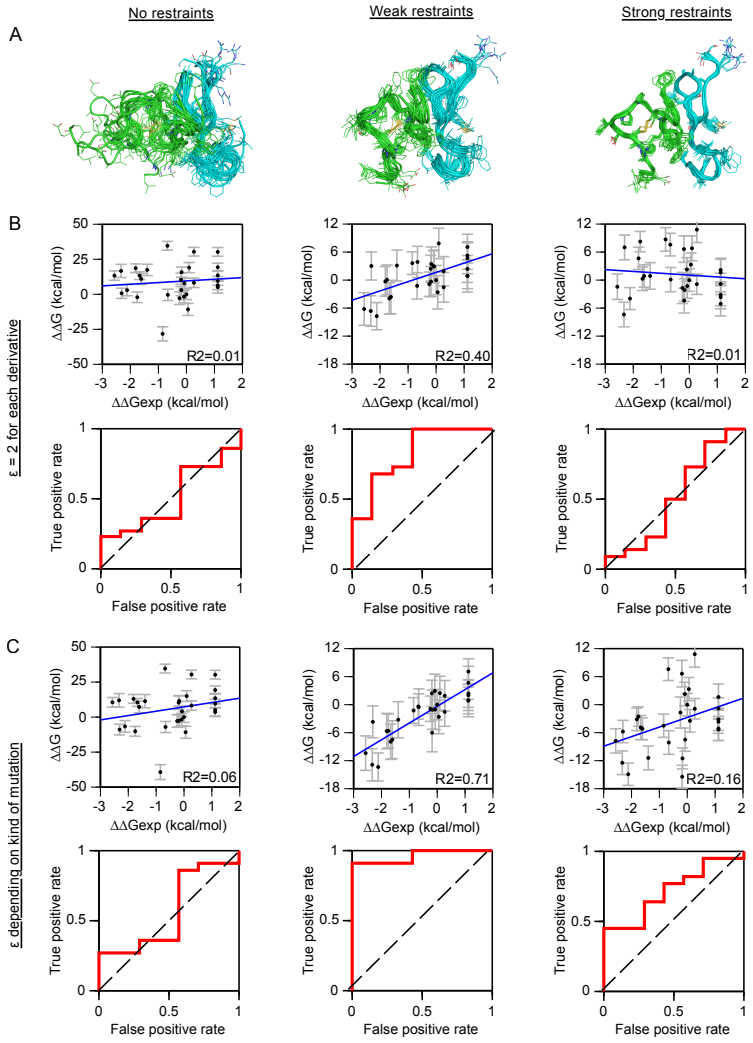


Figure 4.3: **Correlation and predictivity using different restraint conditions on 29 selected VIRIP derivatives.** (A) Representation of three different restraint conditions on VIR-165 (green) and gp41 FP (cyan), ranging from unrestrained (left), weak restraints (middle), and strong restraints (right). In each case, 10 superimposed snapshots of a 10-ns MD simulation are shown. The backbone is shown in ribbon and side chain atoms are represented by wireframe. (B) Correlation (top graphs) and predictivity (bottom graphs showing ROC curves). A correlation is shown of experimental $\Delta\Delta G$ (converted from $IC_{50}$ values) and theoretical $\Delta\Delta G$ (calculated with AMBER using a dielectric constant of $\epsilon_p$ = 2). A line is drawn to visualise the fit. WT-VIRIP was used as reference with theoretical and experimental $\Delta\Delta G$ set to zero. ROC curves are shown at the bottom to demonstrate the prediction of theoretical $\Delta\Delta G$ in each restraint condition. Each graph shows the true positive rate (y-axis) versus the false positive rate (x-axis). The black line indicates the "random selection trend". (C) Same as Figure 4.1B, but theoretical $\Delta\Delta G$ calculated with AMBER using different internal dielectric constant values depending on the peptide sequence (see Table 4.1).

Table 4.1: **List of VIRIP derivatives for comparison with $IC_{50}$ data.**

| Derivative | Sequence[a] | $\epsilon_p$[b] | $IC_{50} \pm SEM$[c] |
|---|---|---|---|
| VIR-344 | LEAIPCSIPPCVFFGKPFVF | 5 | 0.28 ± 0.02 |
| VIR-165 | **LEAIPCSIPPCFAFNKPFVF** | 3 | 0.27 ± 0.04 |
| VIR-345 | LEAIPCSIPPCFLFGKPFVF | 5 | 0.39 ± 0.13 |
| VIR-162 | LEAIPCSIPPCVGFGKPFVF | 5 | 0.73 ± 0.13 |
| VIR-23 | LEAIPMSIPPEVAFNKPFVF | 4 | 4.73 ± 0.61 |
| VIR-148 | LEAIPCSIPPCVAFNKPFVF | 3 | 0.18 ± 0.08 |
| VIR-163 | LEAIPCSIPPCVLFNKPFVF | 3 | 0.84 ± 0.08 |
| VIR-164 | LEAIPCSIPPCVFFNKPFVF | 3 | 0.93 ± 0.05 |
| VIR-175 | LEAIPMSIPPEFLFGKPFVF | 6 | 1.34 ± 0.42 |
| VIR-42 | LEAIPMSIPPEVAFAKPFVF | 6 | 3.45 ± 0.44 |
| VIR-102 | LEAIPMSIPPEVFFNKPFVF | 4 | 0.66 ± 0.06 |
| VIR-18 | LEAIPMSAPPEVKFNKPFVF | 2 | 23.46 ± 0.28 |
| WT-VIRIP | **LEAIPMSIPPEVKFNKPFVF** | 2 | 14.79 ± 2.56 |
| VIR-12 | LEAIPASIPPEVKFNKPFVF | 3 | 13.00 ± 1.04 |
| VIR-21 | LEAIPMSIPPAVKFNKPFVF | 4 | 11.00 ± 4.75 |
| VIR-22 | LEAIPMSIPPEAKFNKPFVF | 2 | 10.64 ± 2.23 |
| VIR-24 | LEAIPMSIPPEVKANKPFVF | 2 | 4.62 ± 1.32 |
| VIR-25 | LEAIPMSIPPEVKFAKPFVF | 4 | 17.41 ± 3.66 |
| VIR-26 | LEAIPMSIPPEVKFNAPFVF | 4 | 10.81 ± 0.68 |
| VIR-13 | LEAIAMSIPPEVKFNKPFVF | 2 | 23.50 ± 5.19 |
| VIR-14 | LEAIPMSIAPEVKFNKPFVF | 2 | 16.33 ± 4.34 |
| VIR-15 | LEAIPMSIPAEVKFNKPFVF | 2 | 9.72 ± 1.66 |
| VIR-27 | LEAIPMSIPPEVKFNKAFVF | 2 | 12.72 ± 10.17 |
| VIR-19 | AEAIPMSIPPEVKFNKPFVF | 2 | >100 |
| VIR-20 | LAAIPMSIPPEVKFNKPFVF | 3 | >100 |
| VIR-39 | LEAAPMSIPPEVKFNKPFVF | 2 | >100 |
| VIR-28 | LEAIPMSIPPEVKFNKPAVF | 2 | >100 |
| VIR-29 | LEAIPMSIPPEVKFNKPFAF | 2 | >100 |
| VIR-30 | LEAIPMSIPPEVKFNKPFVA | 2 | >100 |

(a) Amino acids mutated compared to wild type are underlined. The wild type sequence and VIR-165 as starting NMR structure are shown in bold. (b) $\epsilon_p$ is the internal dielectric constant used for the MM-PBSA calculations. (c) $IC_{50}$ values are given in $\mu$M and were determined previously by Münch *et al.* [1].

weakly restrained simulations, while a random trend is observed for unrestrained or strongly restrained simulations (Figure 4.3B).

### 4.4.3 MM-PBSA simulations: inclusion of multiple internal dielectric constants

The initial results were promising, but the correlation and predictivity could be further improved. Therefore, a third adaptation in addition to restraints and entropy calculations was introduced: the implementation of multiple internal dielectric constant values as a function of the peptide sequence.

Using the procedure outlined in the Materials and Methods section, an improved correlation between the theoretical and experimental $\Delta\Delta G$ values using weak restraints was found. Overall, an $R^2$ value of 0.71 was obtained. As shown in

Figure 4.3C, a clear separation is now achieved between strong and weak binding derivatives. The difference between high and low binding free energy is significant, despite large standard deviations intrinsic to the MM-PBSA methodology. In comparison, the correlation is also present in the initial simulations (where $\epsilon_p$ = 2) in Figure 4.3B, but much weaker. In addition, the correlation is improved in the unrestrained and strong restraint conditions (compare Figure 4.3B,C), though these differences are not significant.

Similar conclusions can be made using ROC-analysis. The sensitivity is improved upon using multiple dielectric constants (Figure 4.3C), since the curve rises instantly to a sensitivity of 0.91. In addition, an area under the curve (AUC) value of 0.96 was obtained, which is considered an excellent predictivity. However, the enrichment is certainly not satisfactory when calculating binding free energy using only one internal dielectric constant. As can be observed from Figure 4.3B, the ROC curve has a high AUC value when using weak restraints, though this is mainly attributed by the specificity (x-axis) and not by the sensitivity (y-axis).

Our results demonstrate that the best predictivity is obtained when all individual components of the binding free energy are calculated. The enrichment disappears when only the molecular mechanics free energy $G_{MM}$ or the solvation free energy $G_{solv}$ is determined. In fact, a random trend is only seen when comparing each individual term of the MM-PBSA method with the experimental binding free energy (see Figure 4.4). $G_{MM}$ has the best predictivity, which is probably a result of the hydrophobic interactions in the complex. A combination of these individual terms ($\Delta G_{sub-tot}$), however, results in an improved prediction of theoretical $\Delta\Delta G$ values. Nevertheless, if ROC curves for binding free energy with configurational entropy included ($G_{tot}$) and excluded ($G_{sub-tot}$) are compared, it is obvious that the enrichment is improved even further when the entropy term is taken into account.

Interestingly, a predictive trend between experimental binding free energy and configurational entropy is observed, although the entropy is predicted as unfavourable in all cases. This is not unexpected, as the entropy components of the complex and the individual peptides were calculated from a single MD trajectory. Therefore, the entropy of the complex is by definition smaller than the sum of its constituents, resulting in an unfavourable entropy contribution upon binding. Furthermore, the configurational entropy values calculated here mainly represent the motions of the side chain atoms, as the backbone atoms were restrained using weak or strong restraints. In fact, the average entropy contribution using weak restraints is 4.48 $\pm$ 2.16 kcal mol$^{-1}$, while a highly similar value of 4.42 $\pm$ 1.41 kcal mol$^{-1}$ was obtained using strong restraints. Thus, it seems that the implementation of restraints does not influence the calculation of the entropic

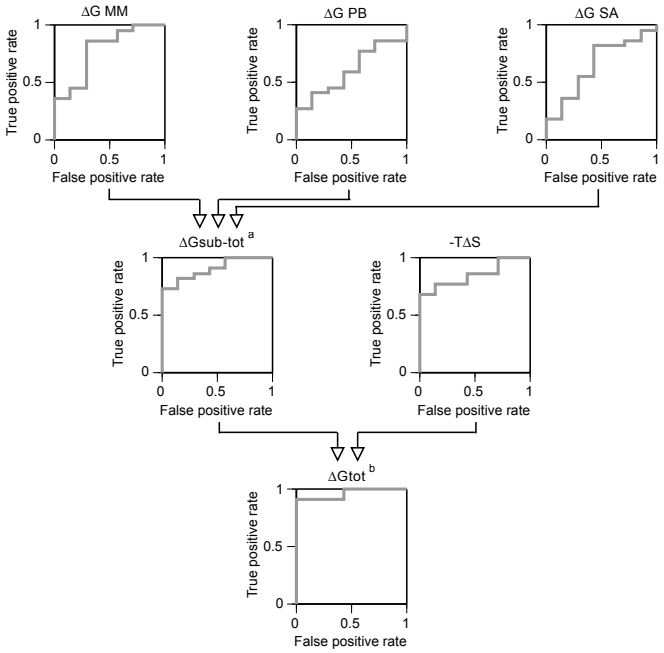


Figure 4.4: **ROC curves for individual binding free energy components.** Weak restraints and multiple internal dielectric constants depending on peptide sequence were applied. Each graph shows the true positive rate (y-axis) versus the false positive rate (x-axis). The black line indicates the "random selection trend". (a) The subtotal binding free energy (excluding entropy): $\Delta G_{sub-tot} = \Delta G_{MM} + \Delta G_{PB} + \Delta G_{SA}$ (b) The total binding free energy (including entropy): $\Delta G_{tot} = \Delta G_{sub-tot} - T\Delta S$.

contributions significantly. Since restraints on the backbone were employed, the configurational entropy contribution should be considered as a virtual screening value instead of an absolute thermodynamic value.

It can be observed from the calculations that the entropy contribution to the total binding free energy is rather small compared to the other individual terms. For example, most optimised VIRIP derivatives have an entropy contribution smaller than 3 kcal mol$^{-1}$. In fact, it has been postulated previously that the optimised VIRIP derivatives have relatively low entropy contributions for binding to FP [1, 21]. In contrast, larger entropic penalties of more than 6 kcal mol$^{-1}$ were calculated for unoptimised VIRIP derivatives. Thus, although the calculated configurational entropy contributions do not significantly improve the predictivity of the setup due to the small magnitude of the entropy values, a relation with the sequence of the peptides was found and its inclusion results in the best enrichment. Possibly, a higher correlation and predictivity could be obtained if translational

and rotational entropy contributions are calculated as well, but these could be difficult to calculate for flexible peptides like VIRIP and FP.

### 4.4.4 Investigation of the VIR-165:FP binding interaction

The construction of a reliable setup for the quantitative study of peptide complexes allows the investigation of the binding interaction of VIR-165:FP in detail. Therefore, an *in silico* computational alanine scan was performed on VIR-165. This procedure allows the identification of essential residues in the binding site, and is based on the assumption that single alanine mutations do not considerably alter the conformation of the peptide complex. As shown in Figure 4.5B, most alanine mutations of VIR-165 are unfavourable. This is a consistent result, since VIR-165 is already an optimised structure compared to wild type VIRIP [1]. Some derivatives show an increased binding free energy, such as L1A and V19A, but the difference is only marginal. Alanine mutations of residues Ile8 and Phe12 have the largest disturbing effect, suggesting that these could be important residues for the binding interaction with FP.

Interestingly, the contribution of each individual amino acid to the binding free energy can be investigated as well, excluding the configurational entropy. The MM-GBSA approach implemented in the AMBER package [10] was used for these calculations, as MM-PBSA cannot be used to decompose the different energy terms. This does not pose a problem as the correlation between MM-PBSA and MM-GBSA values was 0.987 using the testset of 29 VIRIP derivatives (data not shown). From the binding free energy decomposition calculations (Figure 4.5C), it can be deduced that the C-terminal residues of VIR-165 contribute the most to the binding interaction with gp41 FP, where hydrophobic phenylalanine (Phe12, Phe14, Phe18, and Phe20) and a polar asparagine residue (Asn15) are the strongest binding contributors. On the contrary, most N-terminal residues contribute unfavourably to the binding affinity. The largest energy values in the gp41 FP result from a combination of hydrophobic residues (Gly10, Phe11, and Leu12) and polar residues (Ser17 and Thr18). As expected, these residues are positioned closely to the C-terminal residues of VIR-165, as is visible in Figure 4.5A. It is also noteworthy from this figure that the interplay between VIR-165 and FP is a combination of central residues with a strong contribution (shown as red sticks) and residues positioned outward with a weaker contribution (shown as yellow sticks).

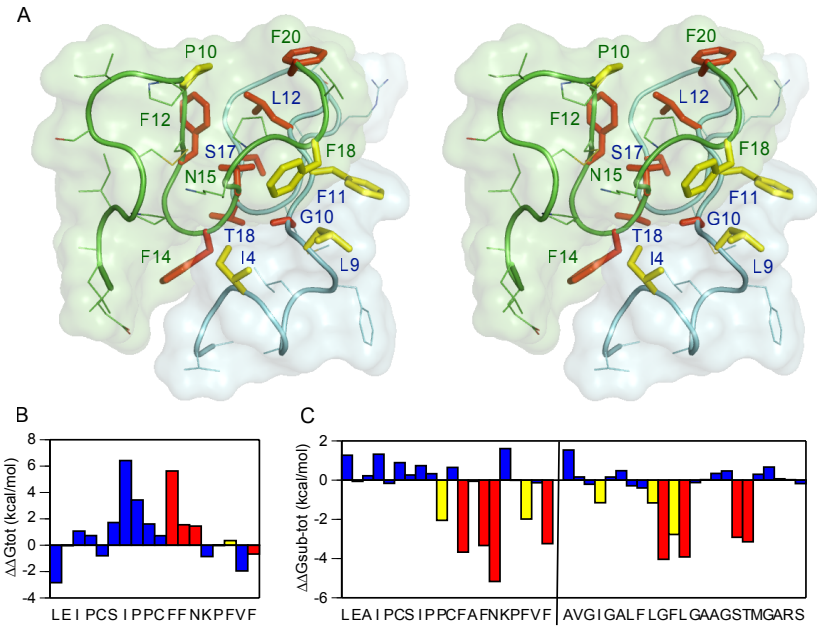


Figure 4.5: **Investigation of the binding interaction between VIR-165 and gp41 FP.** (A) Stereo view of the most important residues of VIR-165 (green) to gp41 FP (cyan) for the binding with ribbon representation and transparent surface. Residues with a binding free energy decomposition below -3 kcal mol$^{-1}$ are shown as red sticks, less important residues between -1 and -3 kcal mol$^{-1}$ are shown as yellow sticks, other residues are shown as lines coloured by element. Main chain atoms and hydrogen atoms were removed for clarity, except for Gly10 of FP that would be invisible otherwise. (B) Computational alanine scan on VIR-165 with calculated free energy difference ($\Delta\Delta G_{tot}$) in kcal mol$^{-1}$. Negative values contribute favourably to the binding, while positive values contribute unfavourably. The sequence of VIR-165 is shown at the bottom. The colour of the bars corresponds to the binding free energy decomposition values of 4.5C. (C) MM-GBSA binding free energy decomposition ($\Delta\Delta G_{sub-tot}$) by residue of VIR-165 and gp41 FP with calculated free energy difference in kcal mol$^{-1}$. Negative values contribute favourably to the binding, while positive values contribute unfavourably. The sequences of VIR-165 and gp41 FP are shown at the bottom. Residues with a large binding free energy contribution (below -3 kcal mol$^{-1}$) are shown as red bars, while less important residues (between -1 and -3 kcal mol$^{-1}$) are shown as yellow bars.

### 4.4.5 Virtual screening

As demonstrated above, the simulation setup yielded promising correlations between theoretical and experimental binding free energies. The structural information and the predictive model were applied in a virtual screening experiment to test the efficacy of the improved MM-PBSA method. Furthermore, this experiment enables the creation of optimised VIRIP derivatives with increased anti-HIV-1 activity. Virtual screening was performed by constructing point mutations in the VIR-165 sequence. A second starting sequence was VIR-175, which lacks the

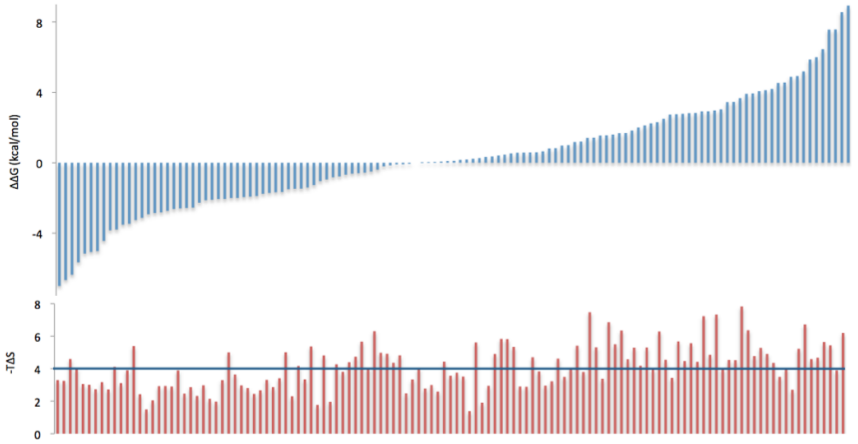


Figure 4.6: **Virtual screening results.** Top: 121 VIRIP derivatives were suggested and ranked by their binding affinity with FP. Bottom: Entropy contribution of each derivative. Due to the importance of the entropy component, we only selected VIRIP derivatives with an entropy threshold below 4 kcal $mol^{-1}$.

disulphide bridge. Although VIR-175 is less active than VIR-165 *in cellulo* [1], it was selected for its therapeutic relevance as it resembles to the most potent molecule currently tested in the clinic, that is, the VIR-576 dimer [1]. The combination of virtual screening on both strong and weak binding derivatives allows the investigation of the predictivity of the MM-PBSA method in a broad affinity spectrum.

In total, 121 mutations were proposed for *in silico* investigation (see Figure 4.6). Twenty derivatives based on VIR-165 and 23 derivatives based on VIR-175 showed improved binding free energy compared to their originating sequences. The majority of the VIR-165 mutations have beneficial binding free energy, while derivatives of VIR-175 have the lowest affinity. All VIR-165 derivatives have disulphide bridges and these additional connections reduce the flexibility of the peptide. As a result, fewer configurations reduce the absolute value of the entropy of each derivative ($-TS$), but the configurational entropy difference upon binding ($-T\Delta S$) is improved because the dynamics of each derivative and the FP are coupled in each simulation. It is interesting to note that the VIR-165 derivatives have the most favourable entropy difference upon binding, as observed by the calculations. These simulations corroborate our initial finding that VIRIP derivatives with disulphide bridges have improved configurational entropy contributions for gp41 FP binding. In addition, replacement of large hydrophobic residues such as phenylalanine and isoleucine is usually unfavourable.

Based on the *in silico* results, 20 specific derivatives were chosen for extensive *in cellulo* analysis: 17 peptides based on VIR-165 and three on VIR-175 were synthesised. Derivatives with beneficial binding free energy were selected. In addition, an entropy limit was imposed on derivatives of VIR-165 because of the importance of this term. It was found that most optimised VIRIP derivatives have configurational entropy contributions lower than 3 kcal $mol^{-1}$. A higher limit of 4 kcal $mol^{-1}$ was used to test this hypothesis and to include VIR-165 derivatives with promising binding free energy values, despite a high entropic penalty. In addition, the reliability of the predictive model was further investigated by also testing a selection of derivatives with decreased *in silico* affinity.

The results of the virtual screening approach are shown in Table 4.2. The *in cellulo* results were obtained by determining the antiviral activity of each VIRIP derivative in HIV-1 inhibition assays as described [1]. The $IC_{50}$ values shown in Table 4.2 were calculated from one representative measurement with X4-tropic HIV-1 NL4-3 and were confirmed in three other independent experiments. No inactive compounds were found in all measurements, thus no substitutions were detrimental for the binding strength. In a separate measurement, cells were infected with R5-tropic HIV-1 virus and the relative inhibitory value differences are comparable in most cases, which is not unexpected as the inhibitory efficacy of VIRIP has been shown to be coreceptor independent [1]. As shown in Figure 4.7, derivatives with significantly enhanced antiviral potency were found. Compared to the $IC_{50}$ of the original sequence using X4-tropic virus, 11 derivatives showed more unfavourable $IC_{50}$ values, two derivatives showed similar inhibitory values, while seven derivatives showed improved efficiency. Some mutations, for example, VIR-165_F12W and VIR-165_F12L_A13F, improved their inhibitory strength by more than twofold compared to the reference VIR-165. Mutations at positions 12 and 13 are very beneficial, as shown by the improved $IC_{50}$ values of VIR-165_F12L_A13F, VIR-165_A13F and VIR-165_F12W. Interestingly, when a computational alanine scan is executed on VIR-165, the mutation of Phe12 has a large disturbing effect (Figure 4.5A). In addition, Phe12 is very important for the binding interaction as shown by binding free energy decomposition of the VIR-165:FP complex (Figure 4.5B). From the *in silico* and *in cellulo* results, it can be concluded that increasing the hydrophobic content in this area is highly advantageous. Other hydrophobic residue substitutions, that is, VIR-165_F14W and VIR-165_I8Y_F12V_A13F, were also favourable. Intriguing derivatives are VIR-175_M6L and VIR-165_S7Q, which have increased affinity although their mutated residues do not directly contact the FP. These substitutions could have a stabilising influence on vicinal amino acids or could stabilise the overall conformation of the peptide. For example, the larger Gln7 residue in VIR-165_S7Q is directed toward Leu1, and this shielding could enhance the hydrophobic interactions of the latter residue with gp41 FP.

Table 4.2: **Virtual screening on selected variants of VIR-165 and VIR-175**[c]

| Variant | $\Delta G_{coul}$ | $\Delta G_{vdw}$ | $\Delta G_{SA}$ | $\Delta G_{PB}$ | $\Delta G_{sub-tot}$ | $-T\Delta S^a$ | $\Delta G_{tot} \pm STDV^a$ | $IC_{50}X4^d$ | $IC_{50}R5^d$ |
|---|---|---|---|---|---|---|---|---|---|
| VIR-165_V19T | -9.61 | -51.12 | -7.60 | 23.68 | -44.65 | 3.30 | -41.35 ± 3.15 | 0.83 | 0.51 |
| VIR-165_I8Q | -14.64 | -50.12 | -7.04 | 27.52 | -44.28 | 3.26 | -41.02 ± 3.35 | 1.03 | 0.80 |
| VIR-165_I4Y | -13.02 | -48.64 | -7.09 | 26.19 | -42.57 | 3.05 | -39.52 ± **0.80** | 1.10 | 0.30 |
| VIR-165_I4L | -11.58 | -49.22 | -7.06 | 25.42 | -42.43 | 3.01 | -39.43 ± **0.87** | 0.62 | 2.03 |
| VIR-165_I8L | -7.99 | -49.54 | -7.11 | 22.67 | -41.97 | 3.17 | -38.80 ± **1.44** | 0.35 | 0.83 |
| VIR-165_A13F | -9.20 | -46.55 | -6.52 | 21.36 | -40.92 | **2.72** | -38.19 ± 2.07 | 0.22 | 0.46 |
| VIR-165_F18L | -10.11 | -45.84 | -6.84 | 21.80 | -40.99 | 3.11 | -37.88 ± 2.92 | 0.48 | 0.40 |
| VIR-165_F14W | -9.42 | -44.68 | -6.66 | 21.99 | -38.77 | **1.49** | -37.28 ± **1.03** | 0.33 | 2.51 |
| VIR-165_I4F | -6.88 | -46.76 | -6.72 | 21.10 | -39.26 | **2.05** | -37.21 ± 2.70 | 0.40 | 0.26 |
| VIR-165_F18Y | -10.07 | -46.86 | -6.52 | 23.36 | -40.09 | **2.93** | -37.17 ± 1.13 | 0.62 | 0.48 |
| VIR-165_F12W | -10.21 | -45.18 | -6.64 | 22.01 | -40.02 | **2.93** | -37.08 ± **1.24** | 0.18 | 1.95 |
| VIR-165_S7Q | -10.01 | -46.78 | -6.73 | 23.63 | -39.88 | **2.90** | -36.98 ± **1.50** | 0.15 | 0.51 |
| VIR-165_F12L_A13F | -9.78 | -43.78 | -6.46 | 20.93 | -39.10 | 3.34 | -35.76 ± 1.19 | 0.13 | 0.41 |
| VIR-165_F12V_A13F_P10A | -12.92 | -48.38 | -7.07 | 23.05 | -45.32 | 4.59 | -40.73 ± **0.31** | 0.61 | nc |
| VIR-165_F12V_A13F_F20Y | -8.30 | -45.87 | -6.66 | 21.76 | -39.08 | **2.81** | -36.28 ± 2.86 | 0.72 | 1.29 |
| VIR-165_I8Y_F12V_A13F | -10.36 | -44.18 | -6.46 | 21.62 | -39.37 | 3.31 | -36.06 ± 3.14 | 0.25 | 0.48 |
| VIR-165_S7N_F12V_A13F | -9.75 | -42.93 | -6.55 | 22.09 | -37.14 | **1.96** | -35.18 ± **1.48** | 0.24 | 0.21 |
| VIR-175_I4Q | -16.38 | -49.82 | -7.84 | 32.69 | -41.35 | 5.00 | -36.35 ± **1.15** | 1.61 | 0.62 |
| VIR-175_M6L | -16.36 | -48.36 | -7.27 | 32.60 | -39.39 | 4.27 | -35.12 ± **1.72** | 0.37 | 0.20 |
| VIR-175_E2Q | -3.45 | -48.93 | -7.01 | 18.33 | -41.05 | 6.31 | -34.74 ± **0.46** | 0.77 | 0.69 |
| VIR-165 | -8.94 | -46.92 | -6.69 | 23.15 | -39.39 | **2.47** | -36.92 ± **1.92** | 0.34 | 0.59 |
| VIR-175 | -14.98 | -47.52 | -7.42 | 30.91 | -39.00 | 6.35 | -32.66 ± 3.36 | 0.47 | 0.20 |

(a) Bold numbers represent favorable entropy values (<3 kcal mol$^{-1}$) or low standard deviations (<2 kcal mol$^{-1}$). (b) nc, not calculated. (c) Binding free energy values given in kcal mol$^{-1}$, $IC_{50}$-values given in $\mu$M. (d) $IC_{50}$-values were determined from both X4- and R5-tropic HIV-1.

Mutations located at the N- and C-terminal part of the VIRIP derivatives were in most cases unfavourable, despite a beneficial *in silico* affinity as predicted by the MM-PBSA method. Evidently, kinetic effects (i.e. docking of both peptides to each other prior to actual binding and folding events) are not estimated by the MM-PBSA approach, which could explain the disagreement between the *in silico* and *in cellulo* values of N- and C-terminal mutations. Although some derivatives based on VIR-165 and VIR-175 with mutations located in the N- and C-terminal part were less active than their references, *in cellulo* alanine scans of the N-terminal or C-terminal residues of wild type VIRIP render the peptide completely inactive [1]. Another important factor is that MM-PBSA is capable of clearly distinguishing active from non-active molecules, that is, predicting the expected trend of the binding free energies, but it performs poorly when derivatives with similar activity are tested [22]. For example, some derivatives that have more unfavourable binding free energies than VIR-165 displayed enhanced activity *in cellulo* (e.g., VIR-165_I8Y_F12V_A13F and VIR-165_F12L_A13F), even though the differences in binding free energies are within range of the standard deviation. In addition, derivatives that are less active than predicted have larger standard deviations regarding the simulated binding free energy (larger than 3 kcal mol$^{-1}$ difference, e.g., VIR-165_V19T and VIR-165_I8Q) and the *in cellulo* effect is consequently more difficult to predict *in silico*. These specific derivatives also have larger configura-

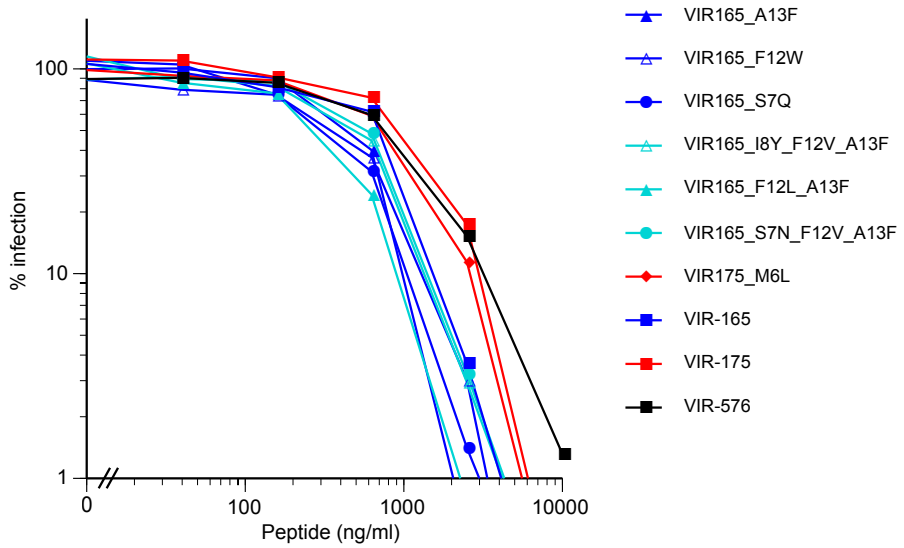


Figure 4.7: **Dose–response curves of improved VIRIP derivatives.** TZM-bl reporter cells were infected with X4-tropic HIV-1 NL4-3. Improved derivatives of VIR-165 are shown in blue, double and triple mutants in teal and improved derivatives of VIR-175 in red. As a remark, VIR-576, shown in black, seems the least active but is a dimer compared to the other derivatives, therefore its $IC_{50}$ value (19 $\mu M$ [1]) is more favorable than visible from this graph. The results were confirmed in four independent experiments.

tional entropy contributions (larger than 3 kcal mol$^{-1}$), but the entropic penalty is still underestimated in these calculations. Thus, although these derivatives have beneficial binding free energy values, the higher entropy contributions explain why these derivatives are less active, highlighting the importance of including configurational entropy calculations for peptide–peptide interactions. In hindsight, an entropy limit lower than 4 kcal mol$^{-1}$ for the selection of VIRIP derivatives should have been imposed, also because derivatives with low simulated entropical contributions are *in cellulo* more active than the original sequence. Although other factors *in cellulo* may influence the reliability of the $IC_{50}$ values, the improved MM-PBSA model has shown to predict several VIRIP derivatives with enhanced affinity *in cellulo*.

Our results show that it is feasible to use computational modelling on peptide complexes to predict mutations that increase *in cellulo* inhibitory activity. The most potent designed derivates have doubled their activity compared to the reference VIR-165. As shown in Figure 4.7, some derivatives have similar or improved potency compared to the derivative currently tested in the clinic, namely, VIR-576 ($IC_{50}$ = 19 $\mu$M)[1]. From our calculations, it can be concluded that peptide opti-

misation will be dependent on reduction of the configurational entropic penalty. Future experiments will include non-natural amino acids to reduce the peptide flexibility and to improve the binding affinity even further. These improved derivatives may initiate small molecule design based on the key features of VIRIP with an analogous mode of action. The improved VIRIP derivative VIR-576 has recently been proven to be safe and effective in a phase I/II clinical trial, but is not suitable for broad application because it requires high doses and it needs to be injected [21]. Peptidomimetics or small molecules with analogous mode of action that are orally available would overcome these limitations. The improved derivatives described in this chapter are a first step toward optimised FP inhibitors for future clinical applications.

## 4.5 Conclusion

MM-PBSA has previously been used for the investigation of small molecule and protein–protein interactions, but we have shown that MM-PBSA can be modified to approximate the binding free energy between peptides. Here, the improved method was applied to investigate the binding interaction of the HIV-1 inhibitor VIRIP with its viral target, the gp41 FP. The best correlation with experimental HIV-1 inhibitory efficiency is obtained when all individual components of the binding free energy are calculated, including the configurational entropy term. Furthermore, the implementation of restraints during the MD simulations was necessary to allow a reliable determination of the binding free energy and to converge the configurational entropy calculations. In addition, an MM-PBSA based virtual screening approach was conducted on two optimised VIRIP derivatives. A selection was tested for antiviral activity and multiple improved derivatives have been identified. Thus, our methodology can form a basis for the development of new peptidic drugs or small molecules with higher affinity.

## References

[1] Münch, J., Ständker, L., Adermann, K., Schulz, A., Schindler, M., Chinnadurai, R., Pöhlmann, S., Chaipan, C., Biet, T., Peters, T., Meyer, B., Wilhelm, D., Lu, H., Jing, W., Jiang, S., Forssmann, W.-G., and Kirchhoff, F. (2007). Discovery and Optimization of a Natural HIV-1 Entry Inhibitor Targeting the gp41 Fusion Peptide. *Cell*. 129(2):263–275.

[2] Delhaise, P., Bardiaux, M., De Maeyer, M., Prevost, M., Vanbelle, D., Donneux, J., Lasters, I., Vancustem, E., Alard, P., and Wodak, S. (1988). The Brugel package: toward computer-aided design of macromolecules. *J. Mol. Graph*. 6(4):219.

[3] Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., and Karplus, M. (1983). CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* 4(2):187–217.

[4] van der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A. E., and Berendsen, H. J. C. (2005). GROMACS: fast, flexible, and free. *J. Comput. Chem.* 26(16):1701–1718.

[5] Jorgensen, W. L. and Tirado-Rives, J. (1988). The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *J. Am. Chem. Soc.* 110(6):1657–1666.

[6] Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., and Klein, M. L. (1983). Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* 79(2):926.

[7] Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A., and Haak, J. R. (1984). Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* 81(8):3684–3690.

[8] Hess, B., Bekker, H., Berendsen, H., and Fraaije, J. (1997). LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.* 18(12):1463–1472.

[9] Darden, T., York, D., and Pedersen, L. (1993). Particle mesh Ewald: An Nlog(N) method for Ewald sums in large systems. *J. Chem. Phys.* 98(12):10089–10092.

[10] Case, D. A., Cheatham, T. E., Darden, T., Gohlke, H., Luo, R., Merz, K. M., Onufriev, A., Simmerling, C., Wang, B., and Woods, R. J. (2005). The Amber biomolecular simulation programs. *J. Comput. Chem.* 26(16):1668–1688.

[11] Onufriev, A., Bashford, D., and Case, D. A. (2004). Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins*. 55(2):383–394.

[12] Basdevant, N., Weinstein, H., and Ceruso, M. (2006). Thermodynamic Basis for Promiscuity and Selectivity in Protein-Protein Interactions: PDZ Domains, a Case Study. *J. Am. Chem. Soc.* 128(39):12766–12777.

[13] Gilson, M. K. and Zhou, H.-X. (2007). Calculation of Protein-Ligand Binding Affinities. *Annu. Rev. Biophys.* 36(1):21–42.

[14] Kongsted, J. and Ryde, U. (2008). An improved method to predict the entropy term with the MM/PBSA approach. *J. Comput. Aided Mol. Des.* 23(2):63–71.

[15] Fogolari, F., Brigo, A., and Molinari, H. (2003). Protocol for MM/PBSA molecular dynamics simulations of proteins. *Biophys. J.* 85(1):159–166.

[16] Foloppe, N. and Hubbard, R. (2006). Towards predictive ligand design with free-energy based computational methods? *Curr. Med. Chem.* 13(29):3583–3608.

[17] Moreira, I. S., Fernandes, P. A., and Ramos, M. J. (2006). Computational alanine scanning mutagenesis—An improved methodological approach. *J. Comput. Chem.* 28(3):644–654.

[18] Petersen, B., Petersen, T. N., Andersen, P., Nielsen, M., and Lundegaard, C. (2009). A generic method for assignment of reliability scores applied to solvent accessibility predictions. *BMC Struct. Biol.* 9:51.

[19] Cheng, Y. and Prusoff, W. H. (1973). Relationship between the inhibition constant (K1) and the concentration of inhibitor which causes 50 per cent inhibition (I50) of an enzymatic reaction. *Biochem. Pharmacol.* 22(23):3099–3108.

[20] Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recogn. Lett.* 27(8):861–874.

[21] Forssmann, W. G., The, Y. H., Stoll, M., Adermann, K., Albrecht, U., Tillmann, H. C., Barlos, K., Busmann, A., Canales-Mayordomo, A., Gimenez-Gallego, G., Hirsch, J., Jimenez-Barbero, J., Meyer-Olson, D., Munch, J., Perez-Castells, J., Standker, L., Kirchhoff, F., and Schmidt, R. E. (2010). Short-Term Monotherapy in HIV-Infected Patients with a Virus Entry Inhibitor Against the gp41 Fusion Peptide. *Sci. Transl. Med.* 2(63):63re3.

[22] Kuhn, B. and Kollman, P. A. (2000). Binding of a Diverse Set of Ligands to Avidin and Streptavidin: An Accurate Quantitative Prediction of Their Relative Affinities by a Combination of Molecular Mechanics and Continuum Solvent Models. *J. Med. Chem.* 43(20):3786–3791.

# Chapter 5

# The flexibility of HIV-1 FP in solution

"Everything Must Converge In Time"

Nick Cave and the Bad Seeds

---

This chapter is an adapted reprint of the article:

Venken, T., Voet, A., De Maeyer, M., De Fabritiis, G., and Sadiq, S. K. (2013). Rapid Conformational Fluctuations of Disordered HIV-1 Fusion Peptide in Solution. *J. Chem. Theory Comput.* 9(7):2870–2874

The original introduction has been shortened to avoid repetition with previous thesis chapters. This article is the result of a combined modelling effort, where dr. K. Sadiq and I contributed equally. We created and conducted the research and wrote the article together, which has been adjusted with helpful suggestions and corrections of the co-authors.

---

## 5.1 Summary

The conformationally flexible HIV-1 FP is indispensable for viral infection of host cells, due to its ability to insert into and tightly couple with phospholipid membranes. There are conflicting reports on the membrane-associated structure of FP and solution structure information is limited, yet such a structure is the target for a novel class of antiretroviral inhibitors. An ensemble of explicit solvent molecular dynamics simulations were initiated from a disordered HIV-1 FP (aggregate time of $\sim$30 $\mu$s). These simulations revealed that while the vast majority of conformations predominantly lack secondary structure, both spontaneous formation and rapid interconversion of local secondary structure elements occurs, highlighting the structural plasticity of the peptide. Therefore, even at this rapid time scale, FP constitutes a diverse and flexible conformational ensemble in solution. Secondary structure clustering reveals that the most prominent ordered elements are $\alpha$- and 3-10-helical subsets of membrane-bound conformations, while trace populations within 2 Å RMSD of all complete membrane-bound conformations are found to

pre-exist in the solution ensemble. Since inhibitor bound conformations of FP are only rarely found, FP inhibitors could function by modulating the conformational ensemble and binding to non-fusogenic FP structures. A thermodynamic characterisation of the most prominent ordered non-fusogenic structures could facilitate the future design of improved FP inhibitors.

## 5.2 Introduction

Most attempts to characterise the FP have been in membrane-bound form (Figures 1.4A-E) [1–5], of which several have been shown stable in membrane simulations [5–9]. Despite this, characterisation of HIV-1 FP in solution may be of crucial importance in understanding the fusogenic process. Furthermore, the recently discovered VIRIP and its optimised derivatives (e.g. VIR-165 and VIR-576) block viral entry by binding to a disordered conformation of FP (Figure 1.4F) [10, 11]. Thus, structural characterisation of the conformational ensemble of FP in solution is of great pharmacological importance if *a priori* predictions of solution structure can be made.

In this chapter, we aim to gain insight into the propensity to form secondary structure elements as well as to qualitatively establish whether any significant structural plasticity of HIV-1 FP is exhibited in solution. Therefore, we used a classical molecular simulation methodology, which we outline more in detail below. We find that none of the membrane bound structures is stable in solution and that FP undertakes rapid fluctuations including $\alpha$-helix and $\beta$-sheet secondary structure elements, i.e. dynamics compatible with a disordered protein.

## 5.3 Materials and Methods

### 5.3.1 Model construction & simulation details

All-atom explicit solvent molecular dynamics simulations using ACEMD [12] on local computer resources were performed starting from each of the conformations A-F shown in Figure 1.4 up to a production time of 300 ns. $\alpha$-helical conformations A-D were extracted from PDBs 1ERF [1], 1P5A [2], 2ARI [3] and 2PJV [4] respectively, taking the first 23 amino acids of FP from the first structure in each PDB. $\beta$-sheet containing conformation E was an adaptation of the fusion peptide based on PDB 3D58 [5], shared by the laboratory of Robert C. Rizzo. Conformation F was extracted from 2JNR [10].

Furthermore, a larger ensemble of simulations (94×300 ns) were performed on the GPUGRID infrastructure [13], starting from a completely unstructured peptide chain, with the highest prevalence HIV-1 FP sequence, AVGIGALFLGFLGAAGST-MGARS, denoted FP23. This structure was built using VMD [14].

The systems were solvated with TIP3P water [15] molecules and neutralised at an ionic concentration of 150 mM NaCl using the leap module of the AMBER 10 software package [16]. All minimisation, equilibration and production simulations were performed using ACEMD [12]. The recently improved AMBER forcefield, ff99SB-ILDN, was used to describe all parameters [17]. Conjugate gradient minimisation was performed for 1000 steps, followed by 1 ns of unconstrained equilibration in the NPT ensemble with a timestep of 4 fs. The temperature was maintained at 300 K using a Langevin thermostat [18] with a damping coefficient of 0.1 and the pressure at 1 bar using a Berendsen barostat [19] with a pressure relaxation time of 800 $ps^{-1}$. The SHAKE algorithm [20] was employed on all atoms covalently bonded to a hydrogen atom. The long range Coulomb interaction was handled using a GPU implementation [21] of the particle mesh Ewald summation method (PME) [22]. A non-bonded cut-off distance of 9 Å was used with a switching distance of 7.5 Å for the VdW interactions.

For the unstructured peptide, the primary phase of equilibration caused collapse of the extended linear chain into a coiled structure thus allowing the construction of a smaller system. The ultimate structure from the primary equilibration was extracted and subsequently resolvated with TIP3P water molecules and ionically neutralised at a concentration of 150 mM NaCl and subjected to a secondary phase of equilibration. Conjugate gradient minimisation was reapplied for 1000 steps and the system requilibrated under identical NPT conditions as before, but with a 2 fs timestep and with a constraining potential of 1 kcal/molA$^2$ on all protein heavy atoms. Unrestrained equilibration was then performed for 10 ns changing the timestep to 4 fs in the NPT ensemble and for a further 10 ns in the NVT ensemble. Different initial structures, obtained from the last phase of the equilibration, were used to seed 100 production simulations, each with randomised velocities describing a Maxwell-Boltzmann distribution. All production simulations were performed in the canonical NVT ensemble on the GPUGRID infrastructure [13]. As several trajectories were not returned by the server and some were more advanced than others at the time of analysis, a subset of 94×300 ns was used for the analysis.

### 5.3.2 DSSP analysis

The DSSP method (define secondary structure of proteins) [23] was used to characterise the secondary structure for each snapshot in each trajectory. The *do_dssp*

plug-in tool in GROMACS (version 4.5.3) [24] was used to implement DSSP. The following secondary structure elements are defined: $\alpha$-helix, $\beta$-sheet, coil, $\beta$-bridge, bend, turn, $\pi$-helix and 3-10-helix for each amino acid.

### 5.3.3 Secondary structural features of membrane-bound structures

The structural features of conformations A-F are defined as: A ($\alpha$-helix: residues 3-15), B ($\alpha/\pi$-helix: residues 4-15), C ($\alpha$-helix: residues 4-18), D ($\alpha$-helix: residues 4-9 and 11-14), E ($\beta$-hairpin: 2-5 and 9-12 and $\alpha$-helix: residues 15-22) and F (bend/turn/coil: all residues). These definitions were used to calculate RMSDs of $C_\alpha$ atoms with respect to the initial conformations with prior fitting to the same secondary structural features, in each case.

## 5.4 Results

### 5.4.1 Solvent MD simulations of experimental membrane structures

Figure 5.1 shows the time evolution of the DSSP state during the 20 ns equilibration and subsequent 300 ns simulation of each of the conformations A-F of FP. All simulations of FP in solvent from each of the conformations A-F show rapid decay of their respective secondary structure features into unstructured conformations within the 20 ns period of equilibration and specifically as soon as constraints are removed. Furthermore, lack of ordered features persists for all systems across 300 ns of simulation, except interestingly for system F. This system reconstituted secondary structural features (specifically a small $\beta$-hairpin) after 100 ns of simulation in a similar sequence region to the initial conformation of model structure E and which persisted for the duration of the simulation.

$C_\alpha$ atom RMSD calculations with respect to the initial secondary structural features of each structure (dark grey lines) are consistent with the DSSP calculations, showing small RMSD values ($< 1$ Å) up to removal of constraints followed by rapid increase (4Å$<$ RMSD $<$ 10Å) upon further equilibration and simulation for all systems. Our results thus show that the respective experimental membrane-bound and VIR-165-bound monomeric apo-form structures are unstable in solution, yet highlight that secondary structural elements may be spontaneously formed,

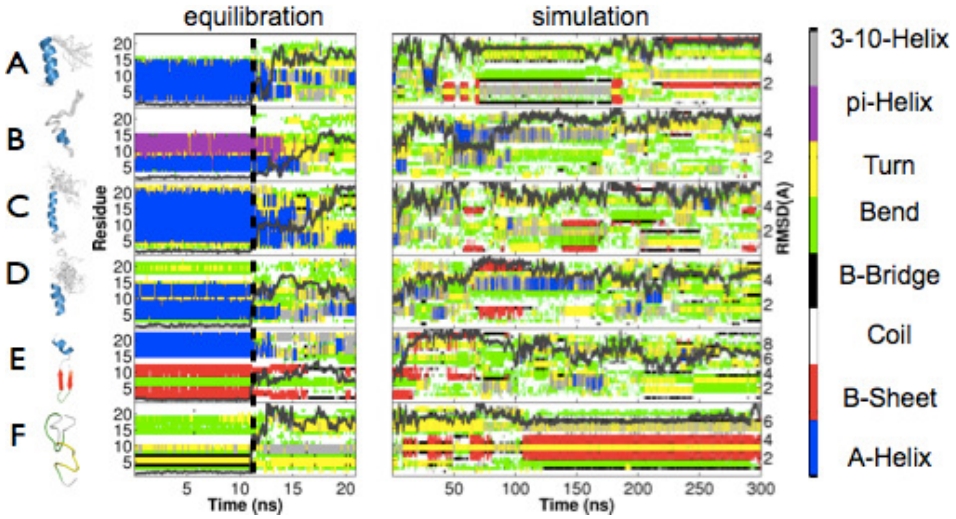


Figure 5.1: **Unfolding of the membrane conformations**. Time evolution of the DSSP state (color legend) and RMSD of $C_\alpha$ atoms of secondary structure regions (dark grey line) is shown during 20 ns equilibration and 300 ns subsequent simulation for HIV-1 FP23 in solution, starting from each of the experimental and modelled membrane-bound structures (A-E) and the VIR-165-bound structure in apo-form (F). Rapid unfolding occurs for each structure as soon as constraints are removed, 11 ns into the equilibration (dashed black line).

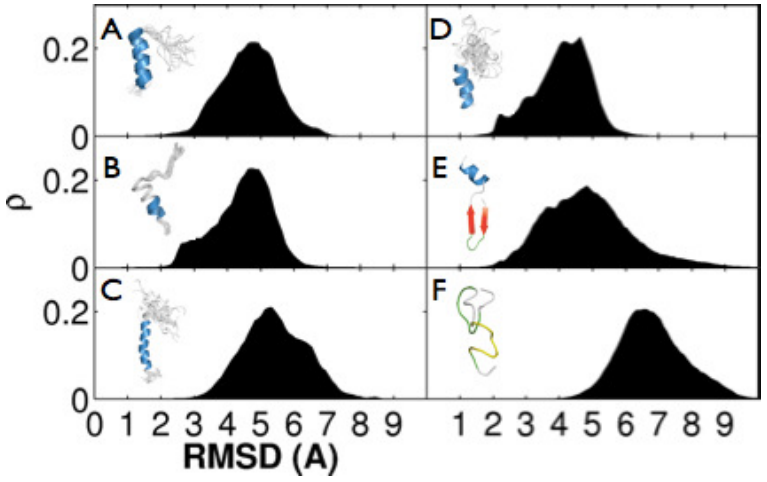


Figure 5.2: **RMSD distribution of the conformational ensemble of FP**. The $C_\alpha$ atom RMSD with respect to corresponding experimental/model FP structures (A-F) was measured from the conformations in the simulation ensemble.

thus further indicating that HIV-1 FP has substantial conformational plasticity in solution.

### 5.4.2 Ensemble solvent MD simulations

We investigated whether any of the structural features found in structures A-F can pre-exist in solution, by performing simulations (94×300 ns) from multiple initial unstructured conformations of FP23.

#### RMSD distribution

We determined the $C_\alpha$ atom RMSD distribution of the entire ensemble with respect to each of the specific secondary structural features corresponding to the experimental conformations (Figure 5.2). Most conformations in our solution ensemble deviate strongly from the FP23 structures found in membrane-like environments (Figure 5.2A-E) with peak RMSDs of between 5 to 6 Å compared to secondary structural features. Nevertheless, a small percentage of structures within the ensemble were found within 2 Å RMSD of conformations A-D (A: 0.21%, B: 0.05%, C: 0.01%, D: 0.49%) while conformation E was not present. Interestingly, when comparing RMSD relative to either the $\alpha$- (residues 15-22) or $\beta$- (residues 2-5 and 9-12) components of structure E separately, 0.21% and 6.38% of conformations were within this RMSD threshold respectively (RMSD distribution for E shown in Figure 5.2 corresponds comparison with the $\beta$-sheet only). The mean deviation from structure F (VIR-165 bound apo-form of FP) was ∼7 Å and no conformations were present below a 2 Å RMSD threshold, the minimum being at 3 Å and only 0.01% were within 3.5 Å RMSD. This is not surprising as conformation F has no significant secondary structural features (only bends and turns), so RMSD was calculated with respect to all $C_\alpha$ atoms within the chain.

#### DSSP proximity distribution

Calculating the proximity distribution in DSSP state space resulted in a similar trend, displaying the rare occurence of structures matching the experimental structures. Proximity of any two snapshots in DSSP space was calculated by computing the number of amino-acid changes required to convert one DSSP structure into the other. A change of DSSP state for a single amino acid into any other DSSP state was considered a distance of 1 in the DSSP space. Of the experimental structures of membrane bound FP (Figure 5.3:A-D), the solution ensemble most closely proximates the $\alpha$-helical conformations A and B, with a

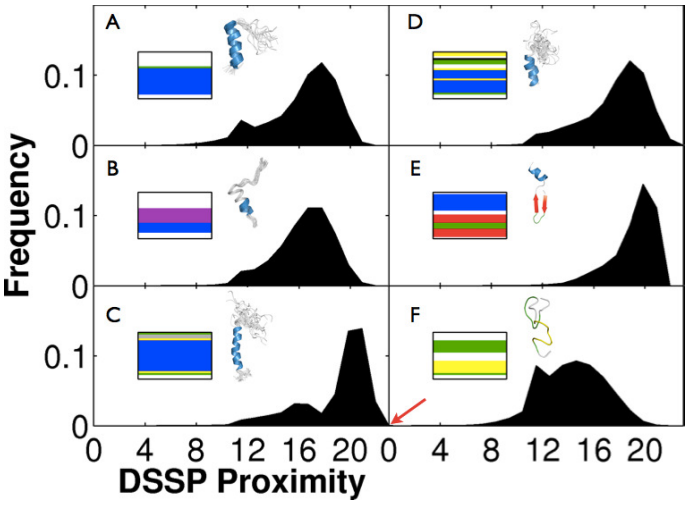


Figure 5.3: **Normalised proximity distribution in DSSP state space for the ensemble with respect to various experimental/model FP structures**. The structures are outlined in Figure 1.4:A-F, respectively. Corresponding DSSP states or 'flags', of these reference structures are inset. The red arrow indicates a single instance in the data set which exactly matches the VIR-165 bound structure of FP.

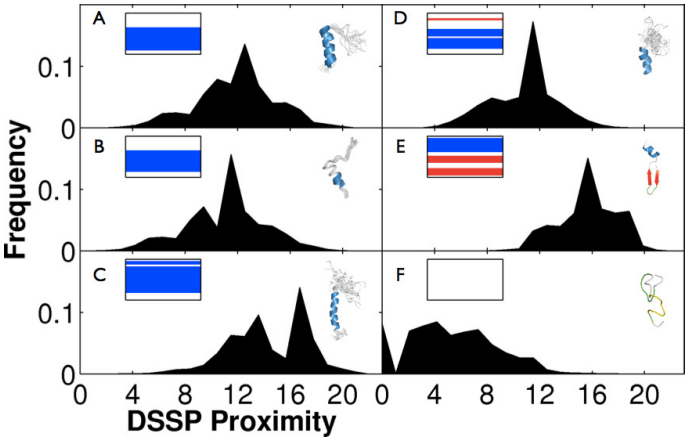


Figure 5.4: **Normalised proximity distribution in reduced DSSP state space for the ensemble with respect to various experimental/model FP structures**. The structures are outlined in Figure 1.4:A-F, respectively. Corresponding DSSP states or 'flags', of these reference structures are inset. In the reduced representation $\pi$-helices and 3-10-helices are included as $\alpha$-helices (blue), $\beta$-bridges are included as $\beta$-sheets (red) and turns and bends are included as coils (white). This allows only the classical secondary structural features to be compared.

peak variation of 16 amino acids in DSSP space and an additional minor peak at a variation of 12 (11 identical amino acids). Thus, although the majority of solution conformations vastly differs from the membrane conformations, some membrane-like structures infrequently seem to pre-exist in solution.

### Reduced DSSP proximity distribution

The proximity distribution of the entire ensemble in reduced DSSP state space with respect to the six experimental/model structures was calculated (Figure 5.4). A reduced representation in which $\pi$-helices and 3-10-helices are included as $\alpha$-helices (blue), $\beta$-bridges are included as $\beta$-sheets (red) and turns and bends are included as coils (white) allows only the classical secondary structural features to be compared. Compared to the full DSSP distribution, there is a notable shift of each distribution towards more proximal states. The key secondary structure features of conformations A (5 instances) and B (14 instances) exist, albeit rarely, in the solution ensemble. Conformations C-E are not found although there is a notable shift in the ensemble distribution to more proximal states. In addition, a small number of coiled structures similar to the VIR-165-bound conformation were found.

### Analysis of conformational changes

Our motivation for using DSSP to gain insight into conformational changes stems from the insufficient information that other conventional metrics often yield. This notion is certainly relevant for an extremely flexible system such as the fusion peptide, and furthermore, when a sufficient number of reference structures do not exist. For example, the normalised frequency distribution of the radius of gyration (Figure 5.5A) shows a single peak at about 7.5 Å; no specific structural information about the different conformational features can be gleaned. Even a frequency distribution of the RMSD with respect to both an $\alpha$-helical (structure A in Figure 1.4) and $\beta$-sheeted (structure E in Figure 1.4) structure (Figure 5.5B) shows a single peak. No member of the ensemble exhibits an RMSD with respect to these conformations of less than 5 Å; furthermore, there exist many distinct conformations within an RMSD of 7 Å that are equidistant from both $\alpha$-helix and $\beta$-sheet structures. By contrast, DSSP directly distinguishes the key secondary structural features. The drawback of DSSP is that such characterisation comes at the cost of approximating tertiary conformational structure to secondary structural features. Therefore, DSSP would not be relevant for tertiary structure characterisation of larger proteins but in the smaller peptide limit, it can provide insight.

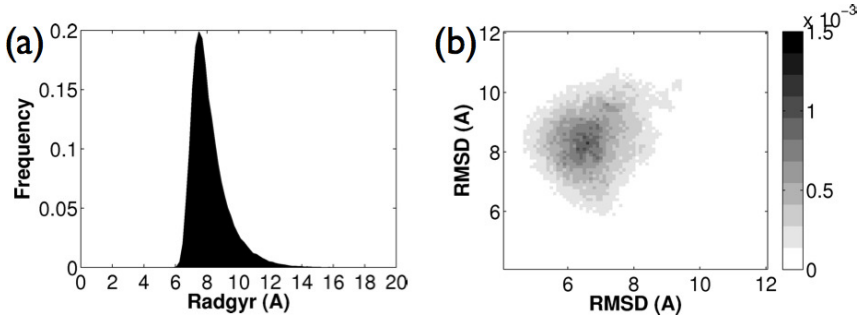


Figure 5.5: **Normalised frequency distributions for the ensemble using radius of gyration and RMSD**. (A) The radius of gyration of the solution ensemble. (B) The RMSD distribution for the ensemble with respect to $\alpha$-structure A (x-axis) and $\beta$-structure E (y-axis) defined in Figure 1.4.

### Global secondary structure propensity

The overall propensity to form any given secondary structure element was calculated both as an average across the entire peptide (termed global DSSP propensity) and on a per residue basis. Block averages of the secondary structures in DSSP representation were calculated in 30 ns intervals to verify equilibrium of the global propensity to form secondary structure elements (Figure 5.6). The global $\alpha$-helix propensity is reduced during the first 100 ns while the $\beta$-sheet and $\beta$-bridge increases gradually. The global amino acid structural properties are in equilibrium after 150 ns. The latter 94$\times$150 ns was therefore used for reporting global and per residue structural propensities.

DSSP analysis shows that FP23 is predominantly unstructured or exhibits locally stabilised structures such as bends and turns in the solution ensemble (Figure 5.7A). Coiled conformations have a high frequency (average of 38.1%) which is mainly manifested in the terminal residues. Moreover, a high concentration of bends (average of 19.7%) and turns (average of 20.6%) are present in the entire structure. Extended secondary structure elements are found as well, though at lower percentages. A modest amount of 3-10-helices (average of 7.9%) was found in the FP23 structure, while the propensity of $\pi$-helix is insignificant (average of 0.1%).

In contrast, a small degree of $\alpha$-helix formation (average of 4.0%) was detected, mostly between residues Gly5 to Leu12. Interestingly, the higher $\alpha$-helical content between residues Gly5 to Leu12 as compared to the other residues has been suggested previously using a secondary structure prediction program (see chapter 4) and is consistent with the results of an $\alpha$-helical prediction algorithm [25] (see Figure 5.8A). There is also a moderate propensity of $\beta$-sheet conformations (average of 5.7%) or isolated $\beta$-bridges (average of 3.9%) in solvent as well. Combining

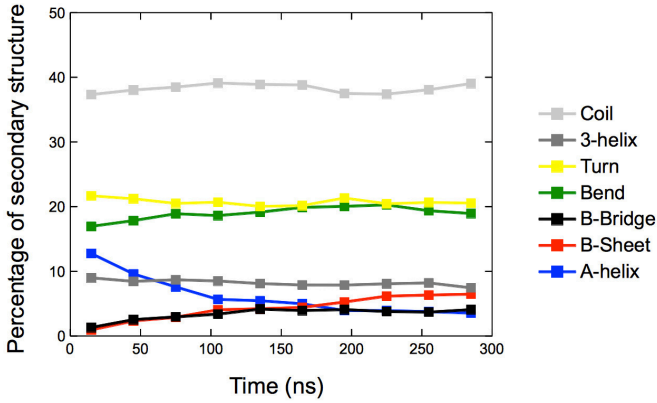


Figure 5.6: **Percentage of average secondary structure of all trajectories**. Block averages were calculated in 30 ns intervals. Structural colour scheme is blue:$\alpha$-helix, red:$\beta$-sheet, white:coil, black:$\beta$-bridge, green:bend, yellow:turn and grey:3-10-helix. The propensity of $\pi$-helix is very low and is therefore omitted for clarity.

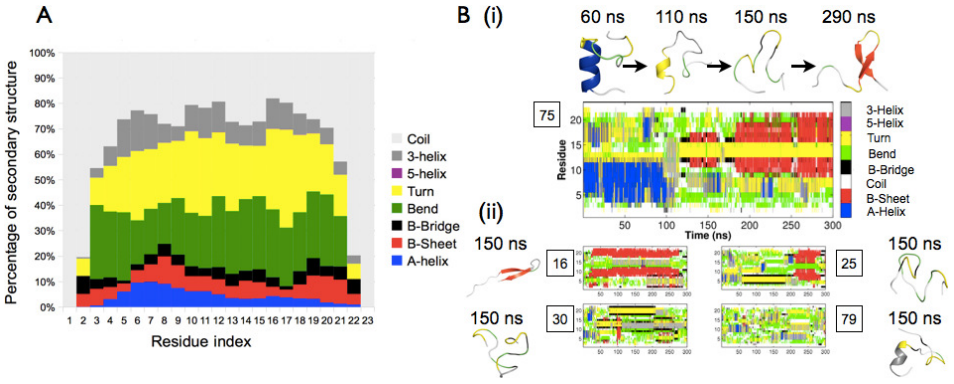


Figure 5.7: **Global secondary structure propensities of the FP ensemble.** (A) Percentage of secondary structure averaged over all trajectories and calculated per residue. Structural colour scheme is blue: $\alpha$-helix, red: $\beta$-sheet, white: coil, black: $\beta$-bridge, green: bend, yellow: turn, purple: $\pi$-helix and grey: 3-10-helix. (B) Selected examples of trajectories with different secondary structure elements in DSSP representation. Simulation 75: conversion of $\alpha$-helical to $\beta$-sheet conformation. Simulation 16: trajectory with a high $\beta$-sheet content. Simulation 25: trajectory with a combination of $\beta$-bridges and $\beta$-sheets. Simulation 30: trajectory with a combination of multiple $\beta$-bridges and 3-helices. Simulation 79: trajectory with a high amount of coiled conformations and only a limited amount of secondary structure. Representative structures during the trajectories are shown besides the DSSP graphs, with the secondary structure of the FP in DSSP colors.

all helical ($\alpha$-helix, 3-10-helix and $\pi$-helix) and $\beta$-sheet conformations ($\beta$-sheet and $\beta$-bridge) results in a slightly larger tendency to form helices (average of 11.9%) than $\beta$-sheet like structures (average of 9.6%). Nevertheless, these fractions are much lower than the unstructured conformations coil, bend and turn taken together (average of 78.4%).

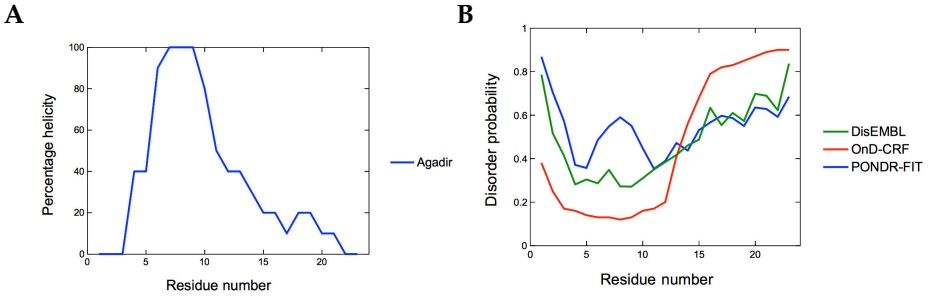


Figure 5.8: **Prediction of secondary structure properties of FP23.** (A) Predicted $\alpha$-helicity of FP23 per residue using AGADIR [25]. This algorithm predicts helicity of peptides based on theoretical helix/coil transitions. (B) Disorder probabilities of FP23 per residue. The PONDR-FIT [26], OnD-CRF [27] and DisEMBL [28] servers (by Loops/coils definition) were used to verify the disorder characteristics of FP23.

### Experimental comparison

We next compared the secondary structure predictions with previous experiments. Transmission FTIR measurements of FP23 in deuterated buffer solution indicated a mixture of molecular aggregates of intermolecular $\beta$-sheet structures and unordered plus helical conformations [29]. FTIR measurements of FP23 in phosphate buffer showed $\alpha$-helix, $\beta$-sheet, $\beta$-bridge and coil estimations of respectively 20.1%, 40.4%, 19.5% and 20% [30]. Electron paramagnetic resonance (EPR) spectra of nitroxide spin labels attached to N-terminal and C-terminal residues showed a high rotation rate, underlining the flexibility of the coiled conformations of FP in solution [31]. These reduced significantly when probing lipid environments. Coiled conformations in aqueous solvent have been found for other fusion peptides as well, such as the influenza FP [32]. In addition, a secondary structure prediction study performed previously in chapter 4 showed $\alpha$-helix, $\beta$-sheet and coil probabilities of respectively 40%, 7% and 53%. Furthermore, several protein disorder prediction methods indicate that the peptide could be disordered (see Figure 5.8B). Although the disorder probability is calculated differently for each method, all three methods suggest that FP23 is mainly disordered. The N- and

C-terminal ends are the most highly disordered, while residues Ile4 to Phe11 display the lowest disorder probability, in qualitative agreement with our results.

The quantitative deviation with our calculations may be for a number of reasons. Firstly, single FP molecules were simulated here while interactions between multiple FP molecules occur in experiments. For example, interstrand $\beta$-sheets and other interresidual stabilisations are absent in our simulations, which could explain the higher $\beta$-sheet content in the FTIR experiments. We hypothesise that conformational experiments on single FPs would result in lower $\beta$-sheet content and partially explain the role of FP multimerisation. Secondly, simulations may be sensitive to the cut-off used to describe various intramolecular interactions [33]. Furthermore, the AMBER forcefield, ff99SB-ILDN, is known to slightly underestimate the $\alpha$-helical content in MD simulations [34]. Use of a more accurate forcefield may thus reduce the observed discrepancy. Finally, while the secondary structure prediction shows a majority of coil probabilities (53%), a large amount of $\alpha$-helix was predicted as well (40%), which is much higher than in our MD simulations and in the FTIR experiments. Secondary structure prediction algorithms are usually optimised for larger proteins which also contain tertiary structures. In contrast, the FP with only 23 amino acids is too small to contain a tertiary fold, which could influence the reliability of the secondary structure prediction.

On the other hand, we note that the amount of $\alpha$-helix measured in experiments may be exaggerated because these involve ensemble measurements averaged in time. Sometimes helical structures can be detected in experiments, though these structures actually could be highly averaged random-flight chains, containing mostly bends, turns and coils instead [35]. Previous studies on the alanine-based XAO peptide, a model for the unfolded state of proteins [36], show it to display a very flexible, fluctuating structure that only seldomly adopts a helical conformation, quite similar to the behaviour of FP23.

### Time evolution analysis

Analysis of the time evolution of secondary structural features across all trajectories reveals a detailed account of the specific conformational ensemble and structural interconversions (Figure 5.9). Strong variations exist between different simulations but also within a single trajectory, clearly demonstrating the inherent structural plasticity of FP23.

Most simulations reveal a high quantity of unstructured conformations. This notion is demonstrated by simulation 79 (Figure 5.7B), for example, where mostly

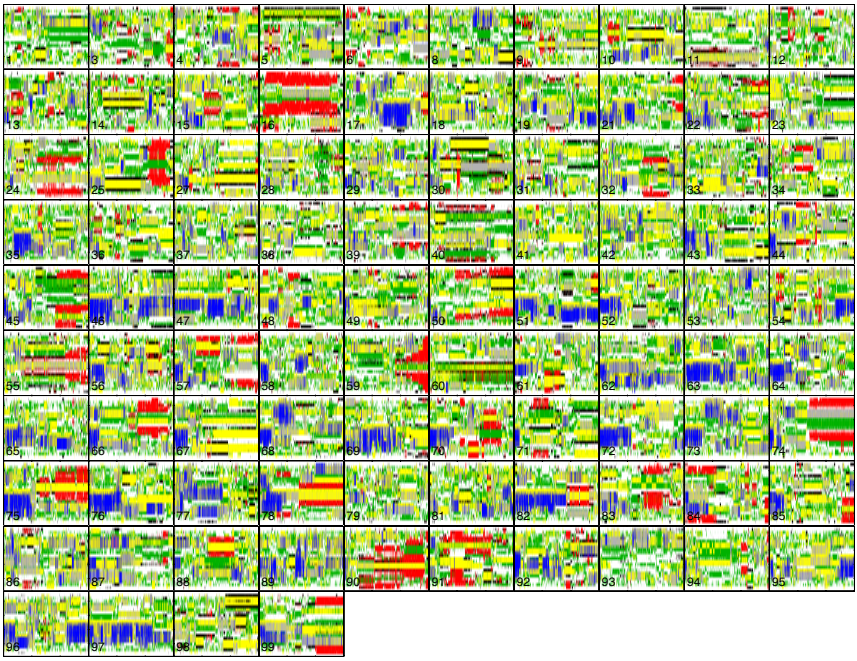

Figure 5.9: **Time evolution of all trajectories in DSSP representation.** Each panel represents a trajectory up to 300 ns. Structural color scheme is blue:$\alpha$-helix, red:$\beta$-sheet, white:coil, black:$\beta$-bridge, green:bend, yellow:turn, purple:$\pi$-helix and grey:3-10-helix.

coil, bend and turn are exhibited. The secondary structures also display a persistence time of only a few nanoseconds. In contrast, simulation 16 clearly illustrates the formation of a stable C-terminal $\beta$-sheet structure, which persists for around 250 ns, though unfolds again at the end of the simulation.

The formation of stable $\beta$-sheets is not restricted to these specific residues as shorter $\beta$-sheets have also been observed, for example, in simulation 75. Isolated $\beta$-bridges were found regularly as well, like in simulation 25. This local structure lasts for more than 100 ns until it unfolds and then forms a transient $\beta$-sheet. $\beta$-bridges can be found in almost all residues except for termini. Simulation 30 for example contains multiple temporary $\beta$-bridges at different time points.

In simulation 75, an $\alpha$-helix with persistence time of $\sim$100 ns converts into a $\beta$-sheet structure that lasts to the end of the simulation via transitional 3-10 helix and $\beta$-bridge conformations (Figure 5.7B and Movie M1[1]). This example clearly demonstrates the high secondary structure plasticity of FP23 in solution, as both

[1]This movie can be found online at http://pubs.acs.org/doi/suppl/10.1021/ct300856r/suppl_file/ct300856r_si_002.mpg

stable $\alpha$-helix, $\beta$-sheet and coiled unfolded structures can occur in just a few hundred nanoseconds.

### Cluster analysis

The ensemble was clustered in DSSP space (Figure 5.10A) using a k-means clustering algorithm in the EMMA software package [37]. A number of cluster sizes were explored with 50 being found to provide a good qualitative balance between differentiating visible features and conformational variance in the ensemble. Clusters contained amino acid regions of both persistent and widely varying ('noisy') DSSP structure, although each cluster was predominantly disordered. Filtering of the most frequent secondary structure elements in each cluster (frequency$>$0.8) and then highlighting those clusters which had low mean RMSDs ($<$0.75Å) relative to the average filtered secondary structure element revealed that the most prevalent and stable conformations in the ensemble are firstly N- and then C-terminal $\alpha$-helices or closely related 3-10-helices. N- and C-terminal $\beta$-sheets also exist in conformations but not to the same degree of prevalence or stability (Figure 5.10B). The most relevant stable elements thus are subsets of the $\alpha$-helical forms of membrane-bound fusion peptide.

## 5.5 Discussion

The conventional paradigm of structure-based function in proteins has been challenged in recent years by the realisation that many proteins exist in a disordered yet functional state [38, 39]. These intrinsically disordered proteins (IDPs) exist in a conformational equilibrium of states with low interconversion energy barriers [40], others become spontaneously ordered in response to triggered binding events [41]. In addition, the entire protein can be disordered or only selected regions can adopt a disordered conformation.

Our results suggest that, on this time scale, HIV-1 FP is mainly disordered and constitutes a diverse, flexible and rapidly interconverting conformational ensemble in solution. However, it remains unclear whether FP23 displays a similar flexibility when attached to the much larger gp41 protein or in its oligomerised state. Crystal structures of gp41 either exclude FP due to its hydrophobicity decreasing solubility and/or because of its flexibility. Nonetheless, we postulate that FP23 as part of a much larger gp41 protein *in vivo* is likely to be disordered based on (a) the disorder exhibited in free FP23 in our study, (b) the fact that several crystal structures of gp41 are not resolved even upon inclusion of the first 23 N-terminal residues

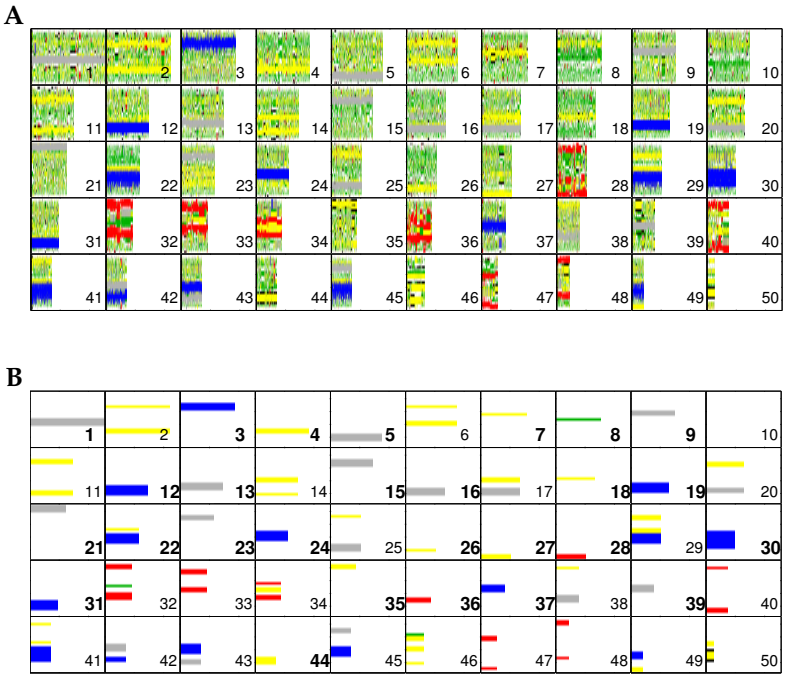


Figure 5.10: **Clustering of the ensemble in the DSSP space**. Clusters are labelled numerically in order of population size within the ensemble. Residue labelling has been removed for clarity. Structural color scheme is blue:$\alpha$-helix, red:$\beta$-sheet, white:coil, black:$\beta$-bridge, green:bend, yellow:turn, purple:$\pi$-helix and grey:3-10-helix. (A) A k-means clustering algorithm with a cluster number of 50 was used. Each panel width is the size of the largest cluster (12813 snapshots). (B) DSSP representations (flags) of the 50 clusters, clustered in the DSSP space. Frequent amino acid regions (frequency$>$0.8) are labelled with their corresponding DSSP state, all others are denoted as coiled. Relative population of each cluster within the whole ensemble is denoted by the length of each flag. Labels in bold correspond to clusters with mean backbone RMSD $< 0.75$ Å relative to the average structure of the corresponding amino acid region in each ensemble.

(for example PDB-codes 2X7R and 3P30), (c) the fact that, to our knowledge, the only crystal structure where a fusion peptide has been resolved, that of Influenza Haemagglutinin in complex with an inhibitor (PDB-code: 3EYJ, chain B), the FP is indeed disordered, and (d) several protein disorder prediction methods indicate that the peptide could be disordered (see Figure 5.8b). In fact, it has been suggested previously that viral fusion peptides may form autonomous folding units in the membrane [32]. We propose that future simulations of gp41 including fusion peptide will be able to test this hypothesis.

Intriguingly, most IDPs are characterised by a high concentration of polar amino acids, yet FP23 lacks these, except for Ser117, Thr118 and Arg122 in the C-terminus.

Nevertheless, FP23 has a high fraction of Ala and Gly residues, which are known for disorder inducing tendencies [42]. The conformational ensemble will evidently be different upon oligomerisation and upon attachment to the entire gp41 protein, which will influence long-range interactions. Nonetheless, the FP23 as such is inserted into the membrane in a monomeric form and is only oligomerised after conformational change of gp41 into a six-helix bundle [43], warranting the analysis of FP23 in its monomeric form to understand membrane insertion.

Recently, an acid-denatured 80 amino acid ACPB protein which has a natural tertiary fold was found to maintain residual ordering and slow interconversions across a long time scale of 200$\mu$s [44]. Our findings cannot rule out the possibility of stable solution structures with relaxation times $>30\mu$s. Such structures would not conform to existing membrane-bound conformations, although they might be composed of a combination of the secondary structural elements identified in our study. However, FP23 is four times smaller than ACPB, lacks a tertiary fold and even structured proteins of slightly larger size fold at significantly shorter time scales [45], suggesting that a stable solution structure of FP23 at longer time scales is unlikely.

## 5.6 Conclusion

Our work opens up future studies that could elucidate the conformational modulation of FP (a) upon formation of a trimer, (b) in the transition from solvent to membrane environment and (c) upon inhibitor binding. The transition from solvent to membrane-bound FP as well as formation of the FP trimer may be crucial in fully understanding the nature of fusogenicity. Importantly, our results show that the key secondary structural features of experimentally observed membrane bound structures are not substantially populated within the solution ensemble due to a large diversity of disordered states, although they partially overlap with the most prevalent of the small minority of ordered conformations. A full kinetic treatment of conformational change upon membrane association, analogous to kinetic studies of fast-folding proteins [46, 47], protein autocatalysis [48] and enzyme inhibitor binding [49] would, in principle, determine the degree to which both conformational selection and induced fit account for FP membrane-insertion and stabilisation. Large-scale simulation of several FPs may determine whether FP readily oligomerises in solution and whether specific conformations facilitate this process.

Our findings suggest that the FP inhibitor VIRIP, which binds a non-fusogenic structure, may function by modulating the conformational ensemble of FP. A

comprehensive view of conformational modulation acknowledges the existence of a distribution of conformations even among the drug-bound states of a protein [40, 50, 51]. The VIRIP bound structure and those that are proximal to it also pre-exist within the solution ensemble, yet occur seldomly. Stabilisation of these structures upon VIRIP binding may shift the conformational equilibrium away from fusogenic structures that co-exist in solution. This would partially explain the high dose requirements of VIRIP in experiments and clinical trials [11] and future investigations may establish this shift quantitatively.

Finally, we suggest an improved strategy to guide the design of new inhibitors. The minority of prevalent ordered elements in the ensemble would be potential targets for inhibition. However, as the most prevalent of these partially overlap with the membrane-bound N-terminal $\alpha$-helical structures, inhibitors targeting such structures would primarily need to disrupt subsequent membrane insertion and/or FP multimerisation. An alternative FP inhibitor design strategy may thus be to thermodynamically characterise the most prominent ordered non-fusogenic structures as targets for inhibition.

# References


[1] Gordon, L. M., Mobley, P. W., Pilpa, R., Sherman, M. A., and Waring, A. J. (2002). Conformational mapping of the N-terminal peptide of HIV-1 gp41 in membrane environments using (13)C-enhanced Fourier transform infrared spectroscopy. *Biochim. Biophys. Acta*. 1559(2):96–120.

[2] Gordon, L. M., Mobley, P. W., Lee, W., Eskandari, S., Kaznessis, Y. N., Sherman, M. A., and Waring, A. J. (2004). Conformational mapping of the N-terminal peptide of HIV-1 gp41 in lipid detergent and aqueous environments using 13C-enhanced Fourier transform infrared spectroscopy. *Protein Sci*. 13(4):1012–1030.

[3] Jaroniec, C. P., Kaufman, J. D., Stahl, S. J., Viard, M., Blumenthal, R., Wingfield, P. T., and Bax, A. (2005). Structure and Dynamics of Micelle-Associated Human Immunodeficiency Virus gp41 Fusion Domain. *Biochemistry*. 44(49):16167–16180.

[4] Li, Y. and Tamm, L. K. (2007). Structure and Plasticity of the Human Immunodeficiency Virus gp41 Fusion Domain in Lipid Micelles and Bilayers. *Biophys. J.*. 93(3):876–885.

[5] McGillick, B. E., Balius, T. E., Mukherjee, S., and Rizzo, R. C. (2010). Origins of Resistance to the HIVgp41 Viral Entry Inhibitor T20. *Biochemistry*. 49(17):3575–3592.

[6] Kamath, S. and Wong, T. C. (2002). Membrane structure of the human immunodeficiency virus gp41 fusion domain by molecular dynamics simulation. *Biophys. J.* 83(1):135–143.

[7] Barz, B., Wong, T., and Kosztin, I. (2008). Membrane curvature and surface area per lipid affect the conformation and oligomeric state of HIV-1 fusion peptide: A combined FTIR and MD simulation study. *Biochim. Biophys. Acta*. 1778(4):945–953.

[8] Taylor, A. and Sansom, M. S. P. (2010). Studies on viral fusion peptides: the distribution of lipophilic and electrostatic potential over the peptide determines the angle of insertion into a membrane. *Eur. Biophys. J.* 39(11):1537–1545.

[9] Promsri, S., Ullmann, G. M., and Hannongbua, S. (2012). Molecular dynamics simulation of HIV-1 fusion domain-membrane complexes: Insight into the N-terminal gp41 fusion mechanism. *Biophys. Chem*. 170:9–16.

[10] Münch, J., Ständker, L., Adermann, K., Schulz, A., Schindler, M., Chinnadurai, R., Pöhlmann, S., Chaipan, C., Biet, T., Peters, T., Meyer, B., Wilhelm, D., Lu, H., Jing, W., Jiang, S., Forssmann, W.-G., and Kirchhoff, F. (2007). Discovery and Optimization of a Natural HIV-1 Entry Inhibitor Targeting the gp41 Fusion Peptide. *Cell*. 129(2):263–275.

[11] Forssmann, W. G., The, Y. H., Stoll, M., Adermann, K., Albrecht, U., Tillmann, H. C., Barlos, K., Busmann, A., Canales-Mayordomo, A., Gimenez-Gallego, G., Hirsch, J., Jimenez-Barbero, J., Meyer-Olson, D., Münch, J., Perez-Castells, J., Standker, L., Kirchhoff, F., and Schmidt, R. E. (2010). Short-Term Monotherapy in HIV-Infected Patients with a Virus Entry Inhibitor Against the gp41 Fusion Peptide. *Sci. Transl. Med*. 2(63):63re3.

[12] Harvey, M. J., Giupponi, G., and Fabritiis, G. D. (2009). ACEMD: Accelerating Biomolecular Dynamics in the Microsecond Time Scale. *J. Chem. Theory Comput*. 5(6):1632–1639.

[13] Buch, I., Harvey, M. J., Giorgino, T., Anderson, D. P., and De Fabritiis, G. (2010). High-Throughput All-Atom Molecular Dynamics Simulations Using Distributed Computing. *J. Chem. Inf. Model*. 50(3):397–403.

[14] Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD - Visual molecular dynamics. *J. Mol. Graphics*. 14:33–38.

[15] Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., and Klein, M. L. (1983). Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys*. 79:926–935.

[16] Pearlman, D. A., Case, D. A., Caldwell, J. W., Ross, W. S., Cheatham, T. E., III, DeBolt, S., Ferguson, D., Seibel, G., and Kollman, P. (1995). AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Comput. Phys. Comm*. 91:1–41.

[17] Lindorff-Larsen, K., Piana, S., Palmo, K., Maragakis, P., Klepeis, J. L., Dror, R. O., and Shaw, D. E. (2010). Improved side-chain torsion potentials for the amber ff99sb protein force field. *Proteins*. 78(8):1950–1958.

[18] Adelman, S. A. and Doll, J. D. (1979). Generalized Langevin equation approach for atom-solid-surface scattering: General formulation for classical scattering off harmonic solids. *J. Chem. Phys*. 64(6):2375–2388.

[19] Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A., and Haak, J. R. (1984). Molecular dynamics with coupling to an external bath. *J. Chem. Phys*. 81:3684–3690.

[20] Ryckaert, J. P., Ciccotti, G., and Berendsen, H. J. C. (1977). Numerical integration of the Cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes. *J. Comput. Phys*. 23:327–341.

[21] Harvey, M. and De Fabritiis, G. (2009). An implementation of the smooth particle mesh ewald method on gpu hardware. *J. Chem. Theory Comput*. 5(9):2371–2377.

[22] Essmann, U., Perera, L., Berkowitz, M. L., and Darden, T. (1995). A smooth particle mesh Ewald method. *J. Chem. Phys*. 103:8577–9593.

[23] Kabsch, W. and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*. 22(12):2577–2637.

[24] Hess, B., Kutzner, C., van der Spoel, D., and Lindahl, E. (2008). GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput*. 4(3):435–447.

[25] Lacroix, E., Viguera, A., and Serrano, L. (1998). Elucidating the folding problem of $\alpha$-helices: Local motifs, long-range electrostatics, ionic-strength dependence and prediction of nmr parameters. *J. Mol. Biol*. 284(1):173–191.

[26] Xue, B., Dunbrack, R., Williams, R., Dunker, A., and Uversky, V. (2010). Pondr-fit: a meta-predictor of intrinsically disordered amino acids. *Biochim. Biophys. Acta*. 1804(4):996–1010.

[27] Wang, L. and Sauer, U. (2008). Ond-crf: predicting order and disorder in proteins conditional random fields. *Bioinformatics*. 24(11):1401–1402.

[28] Linding, R., Jensen, L., Diella, F., Bork, P., Gibson, T., and Russell, R. (2003). Protein disorder prediction:: Implications for structural proteomics. *Structure*. 11(11):1453–1459.

[29] Buzón, V., Padrós, E., and Cladera, J. (2005). Interaction of Fusion Peptides from HIV gp41 with Membranes: A Time-Resolved Membrane Binding, Lipid Mixing, and Structural Study. *Biochemistry*. 44(40):13354–13364.

[30] Gordon, L. M., Nisthal, A., Lee, A. B., Eskandari, S., Ruchala, P., Jung, C.-L., Waring, A. J., and Mobley, P. W. (2008). Structural and functional properties of peptides based on the N-terminus of HIV-1 gp41 and the C-terminus of the amyloid-beta protein. *Biochim. Biophys. Acta*. 1778(10):2127–2137.

[31] Lai, A. L., Moorthy, A. E., Li, Y., and Tamm, L. K. (2012). Fusion Activity of HIV gp41 Fusion Domain Is Related to Its Secondary Structure and Depth of Membrane Insertion in a Cholesterol-Dependent Fashion. *J. Mol. Biol*. 418(1-2):3–15.

[32] Han, X., Bushweller, J. H., Cafiso, D. S., and Tamm, L. K. (2001). Membrane structure and fusion-triggering conformational change of the fusion domain from influenza hemagglutinin. *Nat. Struct. Biol*. 8(8):715–720.

[33] Piana, S., Lindorff-Larsen, K., Dirks, R., Salmon, J., Dror, R., and Shaw, D. (2012). Evaluating the effects of cutoffs and treatment of long-range electrostatics in protein folding simulations. *PLoS One*. 7(6):e39918.

[34] Best, R. B. and Hummer, G. (2009). Optimized Molecular Dynamics Force Fields Applied to the Helix-Coil Transition of Polypeptides. *J. Phys. Chem. B*. 113(26):9004–9015.

[35] Zagrovic, B. and Pande, V. S. (2003). Structural correspondence between the $\alpha$-helix and the random-flight chain resolves how unfolded proteins can have native-like properties. *Nat. Struct. Biol*. 10(11):955–961.

[36] Zagrovic, B., Lipfert, J., Sorin, E. J., Millett, I. S., van Gunsteren, W. F., Doniach, S., and Pande, V. S. (2005). Unusual compactness of a polyproline type II structure. *Proc. Natl. Acad. Sci. U. S. A*. 102(33):11698–11703.

[37] Senne, M., Trendelkamp-Schroer, B., Mey, A. S. J. S., Schütte, C., and Noe, F. (2012). EMMA: A Software Package for Markov Model Building and Analysis. *J. Chem. Theory Comput*. 8(7):2223–2238.

[38] Dunker, A. K., Silman, I., Uversky, V. N., and Sussman, J. L. (2008). Function and structure of inherently disordered proteins. *Curr. Opin. Struct. Biol*. 18(6):756–764.

[39] Mészáros, B., Simon, I., and Dosztányi, Z. (2011). The expanding view of protein–protein interactions: complexes involving intrinsically disordered proteins. *Phys. Biol*. 8(3):035003.

[40] Boehr, D. D., Nussinov, R., and Wright, P. E. (2009). The role of dynamic conformational ensembles in biomolecular recognition. *Nat. Chem. Biol*. 5(11):789–796.

[41] Wright, P. and Dyson, H. (1999). Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol*. 293(2):321–331.

[42] Campen, A., Williams, R. M., Brown, C. J., Meng, J., Uversky, V. N., and Dunker, A. K. (2008). TOP-IDP-scale: a new amino acid scale measuring propensity for intrinsic disorder. *Protein Peptide Lett*. 15(9):956–963.

[43] Doms, R. W. and Moore, J. P. (2000). HIV-1 membrane fusion: targets of opportunity. *J. Cell Biol*. 151(2):F9–14.

[44] Lindorff-Larsen, K., Trbovic, N., Maragakis, P., Piana, S., and Shaw, D. E. (2012). Structure and dynamics of an unfolded protein examined by molecular dynamics simulation. *J. Am. Chem. Soc*. 134(8):3787–3791.

[45] Ensign, D. L., Kasson, P. M., and Pande, V. S. (2007). Heterogeneity even at the speed limit of folding: Large-scale molecular dynamics study of a fast-folding variant of the villin headpiece. *J. Mol. Biol*. 374(3):806 – 816.

[46] Noé, F., Schütte, C., Vanden-Eijnden, E., Reich, L., and Weikl, T. R. (2009). Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. *Proc. Natl. Acad. Sci. U. S. A*. 106(45):19011–19016.

[47] Bowman, G. R., Voelz, V. A., and Pande, V. S. (2011). Taming the complexity of protein folding. *Curr. Opin. Struct. Biol*. 21(1):4–11.

[48] Sadiq, S., Noé, F., and De Fabritiis, G. (2012). Kinetic characterization of the critical step in hiv-1 protease maturation. *Proc. Natl. Acad. Sci. U. S. A*. 109(50):20449–20454.

[49] Buch, I., Giorgino, T., and De Fabritiis, G. (2011). Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations. *Proc. Natl. Acad. Sci. U. S. A*. 108(25):10184–10189.

[50] Hilser, V. J. (2010). Biochemistry. an ensemble view of allostery. *Science*. 327(5966):653–654.

[51] Tsai, C. J., MA, B., Sham, Y. Y., Kumar, S., and Nussinov, R. (2001). Structured disorder and conformational selection. *Proteins*. 44:418–427.

# Chapter 6

## Scrutiny of the Rev multimerisation

"I Get The Same Result"

Non Photo-Blue - Pinback

---

This chapter is an adapted reprint of the article:

Venken, T., Daelemans, D., De Maeyer, M., and Voet, A. (2012). Computational investigation of the HIV-1 Rev multimerization using molecular dynamics simulations and binding free energy calculations. *Proteins.* 80(6):1633–1646

The original introduction and materials and methods section have been shortened to avoid repetition with previous thesis chapters. I performed all the simulations, binding free energy calculations and analysis. I wrote the paper with advice and corrections from the co-authors.

---

## 6.1 Summary

The HIV Rev protein mediates the nuclear export of viral mRNA, and is thereby essential for the production of late viral proteins in the replication cycle. Rev forms a large organised multimeric protein-protein complex for proper functioning. Recently, the three-dimensional structures of a Rev dimer and tetramer have been resolved and provide the basis for a thorough structural analysis of the binding interaction. Here, MD and binding free energy calculations were performed to elucidate the forces thriving dimerisation and higher order multimerisation of the Rev protein. It is found that despite the structural differences between each crystal structure, both display a similar behaviour according to our calculations. Our analysis based on a MM-GBSA and a configurational entropy approach demonstrates that the higher order multimerisation site is much weaker than the dimerisation site. In addition, a quantitative hot spot analysis combined with a mutational analysis reveals the most contributing amino acid residues for protein interactions in agreement with experimental results. Additional residues were found in each interface, which are important for the protein interaction.

The investigation of the thermodynamics of the Rev multimerisation interactions performed here could be a further step in the development of novel antiretrovirals using structure based drug design. Furthermore, the variability of the angle between each Rev monomer as measured during the MD simulations suggests a role of the Rev protein in allowing flexibility of the ARM domain to accommodate RNA binding.

## 6.2 Introduction

Recently, two research groups independently solved the crystal structure of the HIV-1 Rev protein [1, 2]. It is unknown whether the affinity differs between these structures and whether the interfaces have a similar binding free energy contribution per residue, which can differ depending on the conformation of each interface. In addition, an angle of 120° was reported between the $\beta$ interface monomers of 3LPH, while a broader angle of 140° is present between the 2X7L monomers. The question why these subunits have different angles remains unanswered. The difference could originate from protein-protein crystal-packing effects, from respectively the point mutations in 3LPH, and/or the presence of the Fab-fragments in 2X7L.

In this chapter, MD simulation, binding free energy calculations using the MM-GBSA approach [3, 4] and configurational entropy predictions [5, 6] were applied to elucidate the structural integrity and dynamics of the HIV-1 Rev protein as well as to address the questions raised above. Not only the dimer and tetramer crystal structure were investigated, but also a hexamer was constructed to verify the stability of a larger Rev complex. MD simulations allow investigation of the mobility of multiple Rev complexes and individual monomers, while binding free energy calculations and entropy predictions reveal valuable information about the interaction strength and the contribution of individual amino acids at the binding interface. Such information can be very useful to initiate a rational drug design approach. For example, binding free energy calculation on the LEDGF/p75—integrase complex [7, 8] revealed the essential amino acid interactions that have been mimicked by rationally designed inhibitors of the protein-protein interaction [9–11]. As such, our results can be advantageous in the quest for rationally designed protein-protein interaction (PPI) inhibitors targeting the Rev multimerisation process.

## 6.3 Materials and methods

### 6.3.1 System preparation

This chapter is based on the information from two recent crystal structures of the HIV-1 Rev protein: the wild type dimeric structure in complex with two Fab-fragments (PDB entry: 2X7L) and a mutant structure consisting of two dimers, together forming a tetrameric complex (PDB entry: 3LPH). The structure 2X7L comprises two Rev chains with residues 9 to 65, as the disordered C-terminus was not resolvable [1]. The initial structure of 2X7L was constructed from the dimeric Rev protein while retaining Fab fragments at each side. To reduce the amount of atoms in the system, the constant domains of the Fab fragment were deleted. Incomplete arginine residues were modelled in MOE (Chemical Computing Group, Montreal, Canada). The N- and C-terminus of each monomer and the Fab fragments were capped with respectively acetyl and methylamide groups to avoid aspecific interactions. For 3LPH, the protein construct for crystallisation, designated Rev70-dimer, excluded the C-terminus and contained two polar mutations disrupting higher order multimerisation (L12S and L60R) [2]. The 3LPH structure contains a few residues with missing side chains due to low electron density. These missing residues were added with MOE as well. Counterions and crystal waters were stripped. Two amino acids of 3LPH were mutated to leucine (S12L and R60L) to generate the wild type sequence. The termini of each monomer were capped as well. Additional mutant structures of 2X7l and 3LPH were constructed with the MOE software package, where the side chain of the mutated residue was placed according to the rotamer conformation with the lowest energy. To study the behaviour of individual monomers, separate monomers of 2X7L and 3LPH were constructed. To further investigate the different interfaces of the Rev monomers, MOE was used to construct a hexamer using the symmetric crystal packing of 3LPH. This way, symmetric crystal copies were created in all dimensions, but only the two copies interacting through a multimeric assembly were kept. Next, the outer three monomers in each copy were removed. As such, the hexamer contained the original tetramer, but with one monomer added to each side.

### 6.3.2 Naming conventions

The structure 3LPH consists of four resolved chains, each with a different length (monomer A: residue 9 to 70, monomer B: residue 9 to 63, monomer C: residue 8 to 65, and monomer D: residue 8 to 64). The tetramer contains three different interfaces: two similar $\alpha$ interfaces (between chains *A* and *B*, and between *C* and *D*)

and one $\beta$ interface (between chains *B* and *C*). As 3LPH is a tetramer comprising chains ABCD, the hexamer model of 3LPH was named as *ZABCDE*, with E and Z a copy of monomer A and D respectively. Consequently, five different interfaces are present in the hexamer: two $\alpha$ interfaces (*AB* and *CD*) and three $\beta$ interfaces (*ZA*, *BC*, and *DE*). Structure 2X7L comprises two Rev chains with residues 9 to 65. The two monomers in 2X7L form a dimeric conformation, corresponding to the 3LPH *BC* interface. Therefore, in this chapter, the Rev monomers of 2X7L are named monomer B and C. The MD simulations will be referred to as 2X7L and 3LPH for the main simulations, while derived MD simulations are 3LPH-hexamer and the separate monomers: 3LPH-monomerA, 3LPH-monomerB, 3LPH-monomerC, 3LPH-monomerD, 2X7L-monomerB, and 2X7L-monomerC. In total 9 simulations on Rev protein structures were performed.

### 6.3.3 MD simulations

MD simulations were performed with the GROMACS package (version 4.5.3) [12]. Each structure was placed in a dodecahedral box with a minimum distance of 8.5 Å from the edge of the box. The system was solvated with TIP3P water molecules and counterions to neutralise the intrinsic positive charge of the Rev protein. Next, two rounds of energy minimisation were performed, first with 50,000 steps of steepest descent, next with 50,000 steps of conjugate gradient. The system was equilibrated using 100 ps of NVT ensemble with the V-rescale thermostat, followed by 100 ps of NPT ensemble with the V-rescale thermostat [13] and the Parrinello-Rahman barostat [14]. The protein atoms were position restrained during equilibration with a force of 1000 kJ $mol^{-1}$ $nm^{-2}$. After equilibration, a full production MD was executed for 60 ns. The first 10 ns of the run were treated as a further equilibration stage, while data analysis was performed in the remaining 50 ns. To avoid aspecific interactions between the termini of the monomers, position restraints with a force of 1000 kJ $mol^{-1}$ $nm^{-2}$ were applied during the MD production stage on the backbone atoms from the revolved N-terminal residues (residue 8 or 9 depending on the chain) up to Asp11 and from Ser61 up to the resolved C-terminal residues (until residue 63 or higher depending on the chain). The AMBER ff99SB force field [15] was used in all simulations and periodic boundary conditions were applied. The van der Waals interaction cut-off was set to 1.4 nm. For calculation of the long-range electrostatic interactions, the PME summation method was used.

### 6.3.4 Binding free energy calculations

The binding free energy of each protein interface was obtained with the AMBER 8 package [16] using the MM-GBSA method [3] based on the Onufriev-Bashford-Case model [4] as described in section 3.4.1.

Although the Rev molecule is highly charged, the binding interface is mainly hydrophobic, therefore we used an internal dielectric constant ($\epsilon_p = 2$) recommended for a moderately charged interface [17].

For comparative purposes, all binding free energy calculations, decomposition and entropy calculations were performed on the unrestrained residues 12 to 60. The average binding free energy was calculated from 2500 snapshots, taken from the last 10 ns of each production MD simulation. The entropy calculations, as well as the decomposition of the amino acid contribution, were performed on the whole 60 ns to obtain converging results (see Figure 6.4).

### 6.3.5 Analysis tools

The interaxial angle between different monomers during the MD simulations is defined by the intersection of two axes formed by the $C_\alpha$ atoms of Arg35 and Arg58 in each monomer. The angles between the axes of separate monomers were measured in each simulation with 100 ps intervals. Hydrogen bonds were detected between acceptor and donor atom with a cut-off radius of 3.5 Å and cut-off angle of 120°. Using these cut-offs the occupancy of hydrogen bonds was calculated during the last 50 ns of the simulation. Protein structure figures were visualised with Pymol [18].

## 6.4 Results

### 6.4.1 Structural comparison of the different crystal structures using MD simulations

The RMSD of the backbone atoms of each Rev protein MD snapshot structure versus the starting structure is calculated after a backbone least square fit. As is visible in Figure 6.1A, a stable RMSD profile is found for all simulations. The 3LPH-hexamer is a larger structure than 3LPH, but both share a similar RMSD behaviour, with average deviations around 0.2 nm. While the dimeric Rev protein complex of 2X7L is smaller than the tetrameric 3LPH structure, a comparable RMSD profile

is observed during the 60 ns trajectory. The mobility of individual monomer conformations was investigated as well with separate simulations. The average RMSD value for individual monomers was calculated from these simulations (two based on 2X7L and four on 3LPH). As indicated in Figure 6.1B, all monomers display a similar behaviour, although the 2X7L monomers have a slightly higher RMSD profile. A representation of the structures before and after 60 ns of MD simulation is depicted in 6.1C for 2X7L and Figure 6.1D for 3LPH.

The flexibility of different parts of the Rev protein was explored by calculating the root mean square fluctuations (RMSF) of the $C_\alpha$ atoms during the last 50 ns of all simulation systems (see Figure 6.2A). A highly symmetric profile was found, with most mobility in the loop areas ranging from amino acid 30 to 42, intermediate flexibility in the ARM and relatively rigid interfaces where separate monomer conformations interact. This pattern corresponds with the presence of secondary structure in the Rev monomers, though the interaction sites are less flexible due to additional amino acid contacts from vicinal Rev monomers. This behaviour was conserved over all simulation systems. As such, the interfaces seem to act as an anchor point while the loop regions containing the ARM are able to explore the vicinity of the protein, for instance to interact with and adapt to viral RNA binding. Evidently, binding with the RRE would reduce the RMSF of the loop regions and it can be hypothesised that an increased rigidity upon binding could stabilise the overall Rev complex. Comparing the RMSF profiles also shows that the absolute RMSF differences varied between the different monomers in each complex, although the relative differences are conserved over all simulation systems.

Simulations on separate monomers were performed and indicate that all individual monomers have a comparable behaviour: the loop regions containing the ARM show the highest flexibility, but the difference with the other protein segments is reduced compared to the simulations of the complexes (data not shown). The effect of Rev oligomerisation was investigated by calculating the average RMSF difference of all protein atoms for all 2X7L and 3LPH monomers in complex and free simulations (see Figure 6.2B). As such, this RMSF difference explains why amino acids have the most reduced mobility upon binding. As expected, smaller amino acids and regions not important for oligomerisation (such as the ARM) have only limited RMSF difference between complex and free form. Larger amino acids in the interaction sites, however, have a notable reduction in flexibility, for example Arg58 (0.18 nm), Arg48 (0.11 nm), Tyr23 (0.09 nm), Phe21 (0.08 nm), and Glu57 (0.08 nm).

Although 2X7L and 3LPH are highly similar when individual monomers are superposed, the dimers have a different interaxial angle between the monomers.

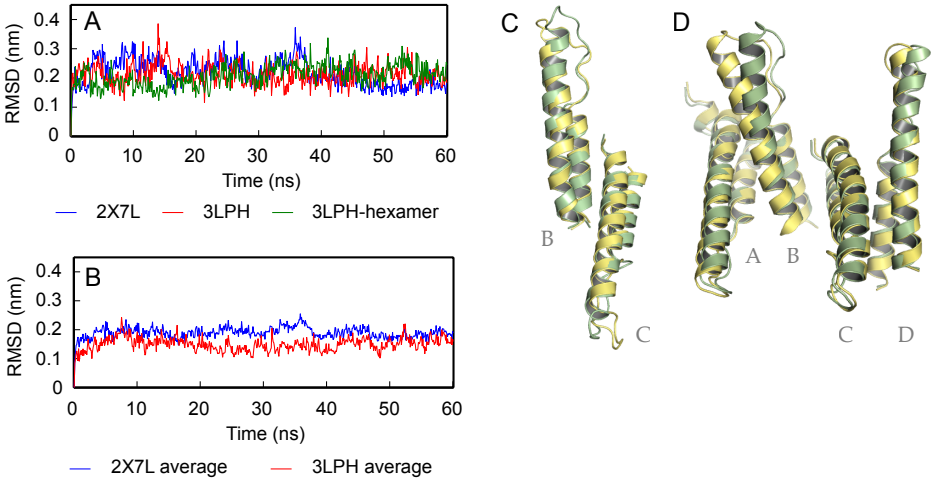


Figure 6.1: **RMSD profile of the backbone atoms of the different Rev protein simulation systems.** (A) RMSD profile of the Rev protein complex simulations. (B) Average RMSD profile per protein structure of independent monomer conformations. (C) Starting structure of the 2X7L dimer (green) and final structure after 60 ns of simulation (yellow). The engineered antibody is omitted for clarity. The chain names of each monomer are shown in grey. (D) Starting structure of the 3LPH tetramer (green) and final structure after 60 ns of simulation (yellow). The chain names of each monomer are shown in grey.

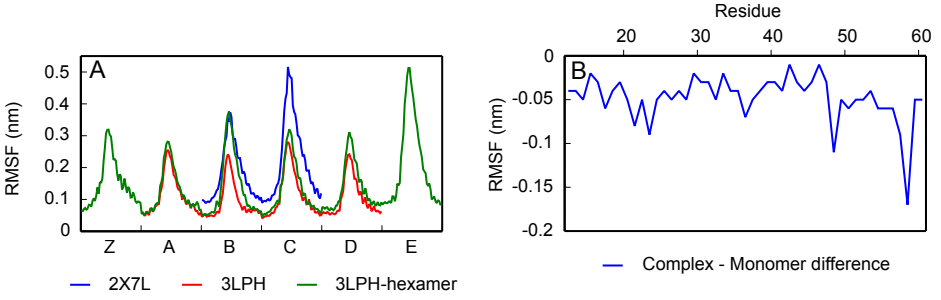


Figure 6.2: **RMSF analysis, showing the flexibility of individual amino acids.** (A) RMSF profile of the $C_\alpha$ atoms of the different Rev protein simulation systems in the last 50 ns of each trajectory. The Rev monomers are chain labeled according to the naming scheme explained in the Materials and Methods section. (B) Difference between the RMSF profile of all protein atoms in the complex structures and the monomeric structures. This RMSF binding difference indicates the flexibility per residue upon complexation of the Rev monomer.

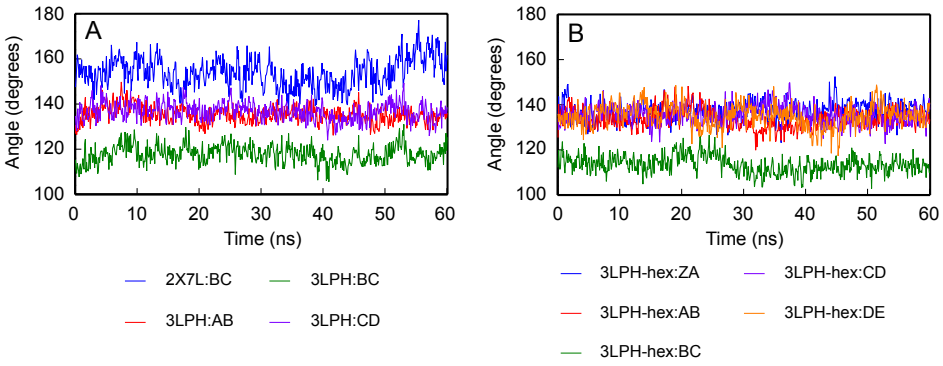


Figure 6.3: **Calculated interaxial angle difference in degrees between the Rev monomers during the 60 ns trajectory with 100 ps intervals.** The C$\alpha$ atom of the top (Arg35) and bottom (Arg58) amino acid of each monomer was used to determine each axis. (A) Angles of the 3LPH and 2X7L systems (B) Angles of the 3LPH-hexamer system.

In the 3LPH structure an angle of 140° was found for the $\alpha$ interfaces [2], while an angle of 120° is present in the $\beta$ interface. In contrast, this angle is 140° in the 2X7L crystal structure [1]. However, it is not clear whether these angle conformations are conserved or simply the result of (i) the mutations in 3LPH, (ii) the different protein constructs, (iii) different crystallisation conditions, or (iv) a combination of previous factors. In Figure 6.3A, the angles between the monomers of 2X7L and 3LPH were calculated from the MD trajectories. During the simulations, fluctuations between roughly 10° and 20° were observed. Interestingly, the angle between monomer A and B and monomer C and D was almost identical in 3LPH. There was, however, a notable angle difference between monomer B and C. This angle was more perpendicular for 3LPH than for 2X7L, resulting in a large angle difference of the $\beta$ interface between both structures. The hexamer of 3LPH was simulated as well and the angles as a function of the MD progress are shown in Figure 6.3B. This structure contains two additional $\beta$ interfaces (*ZA* and *DE*) with an interaxial angle corresponding to the angles found in the $\alpha$ interfaces, but differing from the angle formed by the *BC* monomers. Thus, the presence of two different angle conformations in the 3LPH in the $\beta$ interfaces suggests that the presence of the mutations in the 3LPH protein construct does not explain the angle difference.

Larger fluctuations were observed upon removal of the restraints (data not shown). In fact, permanent transitions were observed in both the unrestrained 3LPH and 3LPH-hexamer simulations, where the interaxial angle increased to the more parallel angle reported in the 2X7L structure. Therefore, different angle conformations can exist in the Rev protein, with anchored interfaces but flexible loop regions

allowing interaction with viral RNA at the ARMs. It should be underlined that these calculations were conducted in the absence of RNA. It can be hypothesised that binding of Rev molecules onto the RRE of viral RNA would stabilise the overall complex and as such would reduce the interaxial angle fluctuations between the Rev monomers. Evidently, the interaxial angle may not be too small *in vivo*, since steric hindrance should be avoided for binding to the RRE.

Overall, the results using RMSD, RMSF, and interaxial angle analysis allow us to conclude that the monomers display a similar behaviour and that no significant difference exists between the crystal structures from a structural point of view.

### 6.4.2 Comparison of the $\alpha$ and $\beta$ interface using binding free energy calculations

It can be hypothesised that the $\alpha$ interface would have more beneficial binding affinity than the $\beta$ interface, since dimerisation occurs before higher order multimerisation [1, 2, 19]. Therefore, binding free energy calculations were conducted using the MM-GBSA method to estimate the affinity between the interaction sites. The results of these calculations are shown in Table 6.1. The binding free energy calculations indicate that the $\beta$ site (*BC* interface) is considerably less favourable than the $\alpha$ site (*AB* and *CD* interface). This does not necessarily imply that each $\beta$ interface binds much weaker than the $\alpha$ interface in all cases. After all, each monomer not only binds to vicinal Rev proteins but also to the RRE motif in a cooperative fashion, and the latter effects have not been taken into account in our calculations. Furthermore, kinetic effects could be present as well [19] but these are omitted from our simulations due to the limitations of the current setup and because these effects are outside the scope of the current research. Interestingly, the binding free energy values of the 3LPH tetramer and hexamer interfaces show a high similarity. The interface of 2X7L has a less favourable binding free energy as compared to the other $\beta$ sites. The difference is not very large though, indicating that the angle differences between the monomers of the 2X7L and the 3LPH simulations do not correlate with the affinity.

### 6.4.3 Apolar contributions

As observed in all simulated systems, the apolar interactions, and more specifically the intermolecular van der Waals forces, are the largest contribution and drive the interaction between the monomers. This hydrophobic nature does not come as a surprise since each binding interface is composed out of a large number of hydrophobic amino acids. Furthermore, most protein-protein interactions are

Table 6.1: **Binding free energy values (given in kcal mol$^{-1}$) calculated with the MM-GBSA method for all Rev simulation systems, with averages of each interface shown in bold.**

| System | $\Delta G_{coul}$ | $\Delta G_{vdw}$ | $\Delta G_{SA}$ | $\Delta G_{GB}$ | $\Delta G_{ele-tot}$ | $\Delta G_{sub-tot}$ | $-T\Delta S$ | $\Delta G_{tot}$ |
|---|---|---|---|---|---|---|---|---|
| $\alpha$ Interfaces | | | | | | | | |
| 3LPH-*AB* | 217.0±7.8 | -86.3±1.4 | -12.6±0.2 | -201.6±7.5 | 15.4±0.4 | -83.5±1.5 | 61.9 | -21.6 |
| 3LPH-*CD* | 237.7±6.9 | -70.5±1.2 | -10.2±0.2 | -222.0±6.6 | 15.7±0.3 | -64.9±1.3 | 58.6 | -6.4 |
| 3LPH-hex-*AB* | 242.4±7.2 | -80.8±1.3 | -11.5±0.2 | -224.0±6.9 | 18.5±0.3 | -73.8±1.4 | 61.7 | -12.1 |
| 3LPH-hex-*CD* | 235.6±6.7 | -70.0±1.2 | -10.1±0.2 | -220.4±6.5 | 15.2±0.3 | -65.0±1.3 | 58.9 | -6.0 |
| **Avg**-*AB*/*CD* | **233.2±11.2** | **-76.6±8.0** | **-11.1±1.2** | **-217.0±10.4** | **16.2±1.5** | **-71.8±7.1** | **60.3±1.8** | **-11.5±7.3** |
| $\beta$ Interfaces | | | | | | | | |
| 3LPH-*BC* | 130.7±4.8 | -39.4±0.9 | -6.4±0.2 | -123.3±4.7 | 7.5±0.2 | -38.4±1.0 | 57.4 | 19.1 |
| 3LPH-hex-*BC* | 148.4±5.3 | -38.8±0.9 | -6.2±0.2 | -140.0±5.1 | 8.4±0.2 | -36.5±1.0 | 58.5 | 22.0 |
| 3LPH-hex-*DE* | 115.5±3.9 | -40.1±0.8 | -6.4±0.1 | -109.3±4.0 | 6.1±0.2 | -40.4±0.9 | 64.2 | 23.8 |
| 3LPH-hex-*ZA* | 110.3±3.9 | -38.6±0.8 | -6.3±0.1 | -104.3±3.8 | 6.0±0.3 | -38.9+0.9 | 64.3 | 25.4 |
| 2X7L-*BC* | 137.4±4.1 | -38.5±0.7 | -6.1±0.2 | -127.2±4.0 | 10.3±0.1 | -34.4±0.8 | 74.0 | 39.6 |
| **Avg**-*BC* | **128.4±15.7** | **-39.1±0.7** | **-6.3±0.1** | **-120.8±14.3** | **7.6±1.8** | **-37.7+2.3** | **63.7±6.6** | **26.0±8.0** |

known to be dominated by hydrophobic interactions, while the introduction of polar interactions are usually disruptive [20]. This effect has been shown by earlier wet lab experiments on the Rev protein [21, 22]. The absolute value of the hydrophobic interactions is much larger for the $\alpha$ interface than for the $\beta$ interface in all simulations. In fact, both the intermolecular van der Waals force $\Delta G_{vdw}$ as the apolar solvation contribution $\Delta G_{SA}$ are almost twice as large for the $\alpha$ sites as compared to the $\beta$ sites. The value of $\Delta G_{SA}$ is proportional to the solvent accessible surface area buried upon complexation, thus the beneficial effect of burying hydrophobic residues when forming complexes is much more favourable for the $\alpha$ interface.

### 6.4.4 Electrostatic contributions

The total electrostatic contribution $\Delta G_{ele-tot}$ is very small compared to the other binding free energy terms such as the apolar contributions. The values are positive and therefore unfavourable for binding, as is the case for many other protein-protein interactions studied before. A closer examination reveals the nature of this unfavourable contribution: the electrostatic contribution in the gas phase $\Delta G_{coul}$ is positive and therefore unfavourable for binding, while the desolvation contribution $\Delta G_{GB}$ is beneficial, though not large enough to obtain an advantageous total electrostatic contribution $\Delta G_{ele-tot}$. This can be explained by the large number of positively charged residues in the Rev protein, mostly in the ARM region. In fact, each Rev monomer has an eight net positive charge. Therefore, the Rev monomers repel each other and this repulsion is too strong to be compensated by the beneficial solvation effects. Evidently, binding to RNA will counterbalance the repulsive effect by neutralising the positive charges. Here, we mainly focus on the Rev oligomerisation domains, but it should be stressed that the presence of RNA could have a strong effect on the situation *in vivo*. However, dimers are

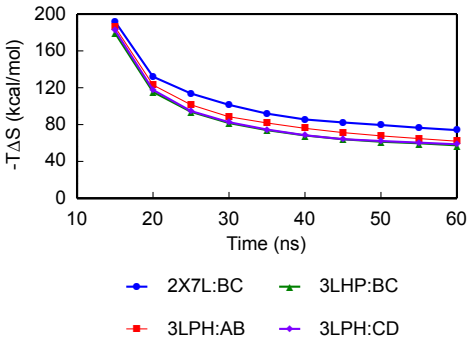


Figure 6.4: **Convergence of the quasi-harmonic approximation of the configurational entropy for each interface in the 2X7L and 3LPH systems.** Data before 10 ns was omitted and considered as equilibration time.

known to form *in vitro*, even in the absence of RNA, as is the case for the crystal structures investigated here. There have been indications as well that Rev is able to form complexes *in vivo* in the absence of RRE RNA [23]. As such, our calculations are still relevant for the study of the protein-protein interactions between the Rev monomers.

### 6.4.5 Entropic contributions

An important binding affinity contribution is the overall entropic contributions between the Rev monomers. As entropy is a time-dependent measure, sufficient sampling is required to obtain convergence of the configurational entropy contribution. Since the entropy calculations did not entirely converge, despite 60 ns of simulation time (see Figure 6.4), the calculations provide only a rough estimate. As seen in Table 6.1, no significant differences were found between the $\alpha$ and $\beta$ interfaces. The entropy contributions are remarkably similar between all interfaces and therefore do not explain the difference in binding between initial dimerisation and higher order multimerisation. This is not unexpected as it has been shown that similar molecules usually do not have distinct configurational entropy contribution, and therefore this involvement is often ignored in MM-GBSA calculations [7, 8, 24–26]. However, it is observed that the 2X7L simulation shows the highest entropic penalty as compared with 3LPH systems, indicating that the dynamics between the monomers are less tightly coupled in the 2X7L system.

### 6.4.6 Hot spot residues at the binding interfaces unraveled by binding free energy decomposition and mutational analysis

Binding free energy decomposition per residue was performed with the MM-GBSA method to identify hot spot residues in each interface and to explain the different binding behaviour between the strong $\alpha$ interface and the weaker $\beta$ interface. To unravel the most crucial amino acids in each binding interface, the binding free energy of all simulation systems was averaged, as shown in Table 6.2. Averaging over all systems ensures that only the most important amino acids in each interface are found. These results are depicted visually in Figure 6.5. Concerning important amino acids, different residues compared to the higher order multimerisation interface guide the initial dimerisation. Specific residues crucial in the $\alpha$ interface barely contribute in the $\beta$ interface, and vice versa. It is notable from Figure 6.5 that the most important residues of the $\alpha$ interface are positioned on both $\alpha$-helices, while the $\beta$ interface interaction is dominated by interactions in the N-terminal $\alpha$-helix. In fact, only one residue (Leu60) in the C-terminal $\alpha$-helix of the Rev monomer makes a significantly strong binding free energy contribution. While Leu64 also exhibited high binding affinity, this residue was not present in all crystallised protein chains and should probably be abandoned as a hot spot.

Comparing respectively the $\alpha$ and $\beta$ interfaces from the 2X7L and 3LPH structures, each interface seems to be conserved and is thus not crystal structure dependent. This is somewhat surprising as two point mutations (L12S and L60R) are present in the 3LPH structure that inhibit multimerisation largely due to electrostatic repulsion. However, it can be postulated that this repulsion is compensated by a stronger hydrophobic effect in the highly saturated conditions of the protein crystal. This is corroborated by the binding free energy calculations shown in Table 6.2. Polar residues in the interfaces like Lys14, Arg17, Ser25, Arg48, and Arg58 are able to represent a favourable binding affinity due to hydrophobic van der Waals interactions that compensate unfavourable electrostatic repulsions. Therefore, it can be concluded that the presence of mutations in 3LPH probably does not influence the overall interaction pattern significantly, despite the crystal-packing process which forces the dimers against each other. Furthermore, the hot spots found in the 3LPH structure were also found in the 2X7L structure, indicating that the complexation with the Fab fragments does not have a disturbing influence on the structural organisation of the 2X7L $\beta$ interface.

The decomposition performed here mainly focuses on enthalpic contributions, but a thorough investigation of the configurational entropic effects upon binding might be necessary as well. Both interfaces show similar entropic binding values, but the entropic penalty could be caused by distinct amino acids in each interface.

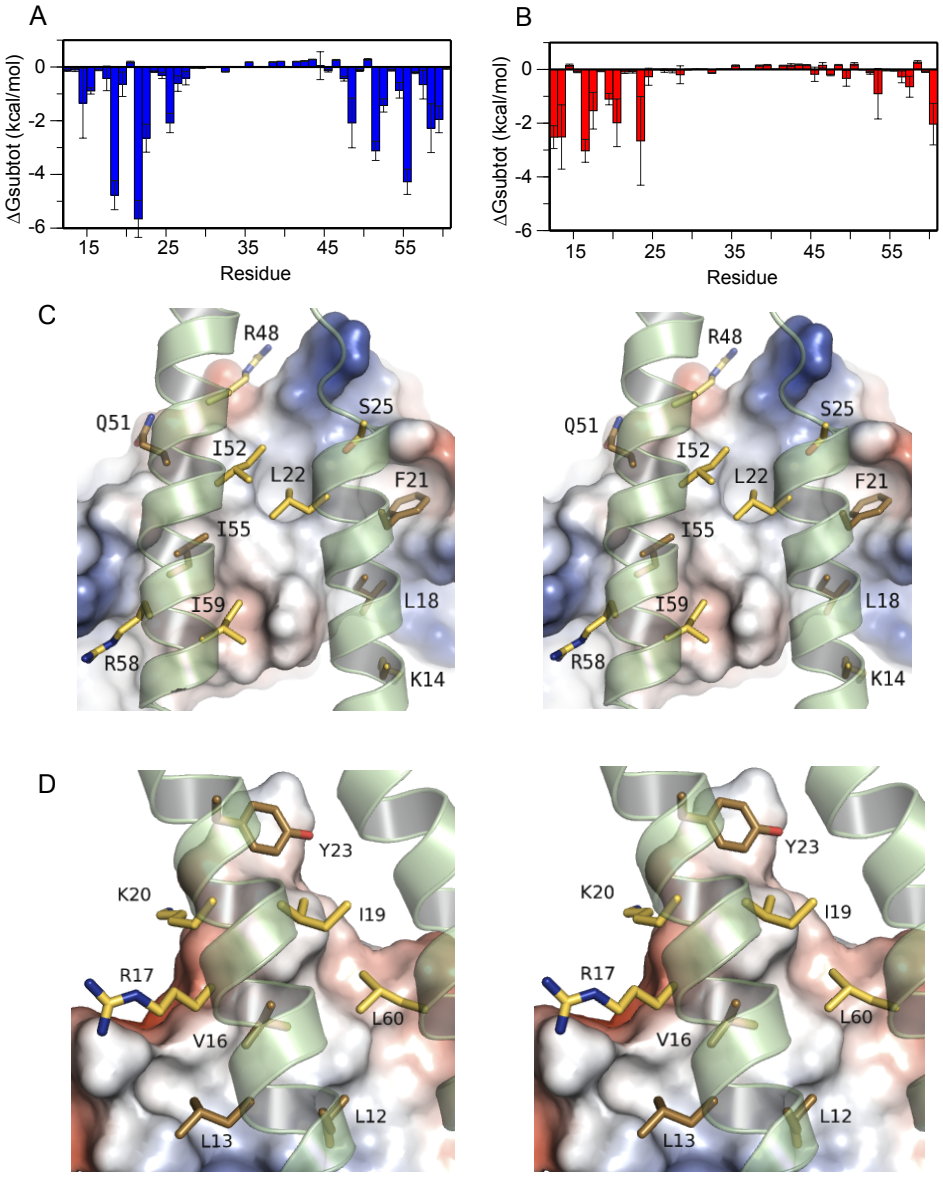


Figure 6.5: **Decomposition of the Rev MM-GBSA binding free energy.** The stereo views of the interactions display one Rev monomer in transparent cartoon representation and the other Rev monomer in electrostatic surface representation. The most important residues in the stereo view are coloured by element in brown, while less important residues are coloured by element in yellow. (A) The $\alpha$ interface per residue, averaged over all simulation systems. (B) The $\beta$ interface per residue, averaged over all simulation systems. (C) Stereo view of the AB interface of 3LPH, as an example of the $\alpha$ interface. D) Stereo view of the BC interface of 2X7L, as an example of the $\beta$ interface.

Table 6.2: **Binding free energy decomposition (given in kcal mol$^{-1}$) calculated with the MM-GBSA method and averaged over all Rev simulation systems.**

| Residue[a] | $\Delta G_{coul}$ | $\Delta G_{vdw}$ | $\Delta G_{SA}$ | $\Delta G_{GB}$ | $\Delta G_{ele-tot}$ | $\Delta G_{sub-tot}$ | $-T\Delta S$ | $\Delta G_{tot}$ |
|---|---|---|---|---|---|---|---|---|
| $\alpha$ Interfaces | | | | | | | | |
| Lys14 | 15.5 | -1.8 | -0.3 | -14.8 | 0.8 | -1.3 | 5.7 | 4.4 ± 0.2 |
| **Leu18**[b,c] | **0.2** | **-4.3** | **-0.6** | **0.0** | **0.2** | **-4.8** | **3.7** | **-1.1 ± 0.4** |
| **Phe21**[b] | **-0.0** | **-5.2** | **-0.9** | **0.4** | **0.4** | **-5.7** | **3.5** | **-2.2 ± 0.5** |
| Leu22[b,c] | 0.1 | -5.2 | -0.3 | 0.1 | 0.2 | -2.7 | 2.8 | 0.2 ± 0.3 |
| Ser25 | -1.2 | -1.6 | -0.4 | 1.2 | 0.0 | -2.1 | 1.4 | -0.7 ± 0.4 |
| Arg48 | 13.5 | -2.5 | -0.5 | -12.6 | 0.9 | -2.1 | 4.3 | 2.3 ± 0.4 |
| **Gln51** | **-0.5** | **-3.4** | **-0.6** | **1.3** | **0.9** | **-3.1** | **2.0** | **-1.1 ± 0.4** |
| Ile52[c] | -0.3 | -1.3 | -0.1 | 0.3 | 0.0 | -1.4 | 2.8 | 1.4 ± 0.2 |
| Ile55[b,c] | -0.2 | -3.8 | -0.6 | 0.3 | 0.1 | -4.3 | 1.6 | -2.7 ± 0.4 |
| Arg58 | 22.4 | -3.0 | -0.6 | -21.1 | 1.3 | -2.3 | 10.0 | 7.7 ± 0.4 |
| Ile59[b,c] | 0.0 | -1.7 | -0.3 | 0.1 | 0.0 | -1.9 | 4.8 | 2.8 ± 0.3 |
| $\beta$ Interfaces | | | | | | | | |
| **Leu12**[b,c,d] | **0.7** | **-2.3** | **-0.4** | **-0.5** | **0.2** | **-2.5** | **2.3** | **-0.1 ± 0.3** |
| **Leu13**[b] | **0.3** | **-2.5** | **-0.4** | **-0.2** | **0.1** | **-2.8** | **1.8** | **-0.9 ± 0.3** |
| **Val16**[b,c,d] | **0.3** | **-2.8** | **-0.5** | **0.0** | **0.2** | **-3.0** | **2.2** | **-0.8 ± 0.4** |
| Arg17 | 0.4 | -2.8 | -0.3 | -1.1 | -0.7 | -1.6 | 5.7 | 4.2 ± 0.5 |
| Ile19[c,d] | 0.3 | -1.1 | -0.1 | -0.1 | 0.1 | -1.1 | 0.7 | -0.4 ± 0.2 |
| Lys20[e] | 2.1 | -2.2 | -0.5 | -1.5 | 0.6 | -2.1 | 5.7 | 3.6 ± 0.5 |
| **Tyr23**[e] | **-0.2** | **-2.4** | **-0.4** | **0.4** | **0.3** | **-2.5** | **3.6** | **1.1 ± 0.4** |
| Leu60[b,c,d] | -0.8 | -1.8 | -0.4 | 0.8 | 0.1 | -2.2 | 3.2 | 1.0 ± 0.3 |

The four most important residues for each interface (around -3 kcal mol$^{-1}$ subtotal binding free energy per residue or better) according to the computational estimations are shown in bold. (a) Only residues with a subtotal binding free energy more favourable than -1 kcal mol$^{-1}$ are shown. (b) Key residues identified by Daugherty *et al.* by buried surface area analysis. (c) Key residues identified by Jain and Belasco by mutational analysis. (d) Key residues identified by Dimattia *et al.* by buried surface area analysis. (e) Residues interacting with the multimerisation inhibiting llama single-domain nanobody ($Nb_{190}$) as shown by Vercruysse *et al.*

Therefore, an entropy decomposition based on the quasi-harmonic approximation was performed [5, 6]. The results are shown in Figure 6.6. Positive numbers imply a large entropic cost, while negative numbers are beneficial to binding. As expected, the configurational entropy correlates well with the RMSF difference values of the amino acids calculated above (see Figure 6.2B). The configurational entropy per residue calculated here allows a thermodynamical estimation of the amount of restructuring of the amino acid side chains upon binding. This entropy can be added to the subtotal binding free energy $\Delta G_{sub-tot}$ to obtain a total binding free energy per residue. Usually, entropic contributions are overestimated using quasi-harmonic analysis, so the magnitude of the entropy values might differ from the actual thermodynamic picture. As a remark, only unrestrained amino acids were investigated (Leu12 until Arg60 in each chain), so it cannot be excluded that amino acids outside this region could play a role in the binding of adjacent monomers. However, most of these amino acids are positioned away from the binding sites and therefore they can be neglected.

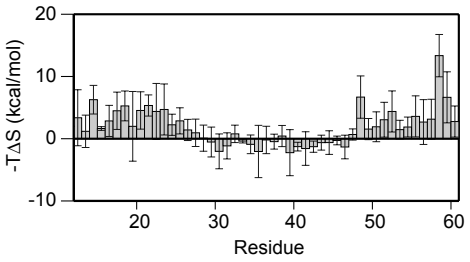


Figure 6.6: **Decomposition of the Rev configurational entropy** The decomposition is averaged over all simulation systems.

### 6.4.7 Hot spot residues in the dimerisation interface

For the $\alpha$ interface, Leu18, Phe21, Gln51, and Ile55 show the strongest binding free energy, both when excluding and including configurational entropy. These four residues form a central interaction core and are positioned at the intersection of the N- and C-terminal $\alpha$-helix. Leu18, Phe21, and Ile55 have been identified previously as essential residues by experimental methods [1, 2, 21]. In addition to these already reported hot spots, Lys14, Ser25, Arg48, Gln51, and Arg58 stabilise the Rev dimerisation interface. In this interface, Arg58 has the largest entropic penalty, indicating that the inherent flexibility of this amino acid might counteract the binding event. Other notable residues with a strong entropic penalty upon binding are Lys14, Leu18, Phe21, Leu22, Arg48, Ile52, and Ile59. In contrast, some residues have an advantageous entropic difference, but the differences are in all cases only small and therefore not significant.

### 6.4.8 Hot spot residues in the higher order multimerisation interface

In the $\beta$ interface, mainly Leu12, Leu13, Val16, Lys20, Tyr23, and Leu60 contribute to the binding between the monomers. Arg17 is another key residue and has not been identified as such before, although the subtotal binding free energy might be counteracted by an entropic penalty. As such, the two point mutations that were introduced to generate the crystal structure of 3LPH (L12S and L60R) seem to play a crucial role in the oligomerisation of the Rev protein from an affinity point of view. However, from our calculation it can be concluded that these mutations do not contribute at all to the affinity of the initial dimerisation site. Thus, the results support the notion that these mutations allow the formation of dimers but not of higher order complexes, which is in line with experimental evidence [21, 22]. In

Table 6.3: **Relative binding free energy differences (given in kcal mol$^{-1}$) of the mutated Rev structures compared to the corresponding wild type structures.**

| Complex[a] | $\Delta\Delta G_{coul}$ | $\Delta\Delta G_{vdw}$ | $\Delta\Delta G_{SA}$ | $\Delta\Delta G_{GB}$ | $\Delta\Delta G_{ele-tot}$ | $\Delta\Delta G_{sub-tot}$ |
|---|---|---|---|---|---|---|
| $\alpha$ Interfaces | | | | | | |
| 3LPH-L18T | 15.3 | -1.9 | -0.3 | -13.7 | 1.6 | -0.6±1.6 |
| 3LPH-I55N | 16.6 | 10.5 | 0.4 | -14.5 | 2.1 | 13.0±1.4 |
| $\beta$ Interfaces | | | | | | |
| 2X7L-L12S/L60R | 42.0 | 15.1 | 1.6 | -40.5 | 1.5 | 18.1±0.9 |
| 3LPH-L12S/L60R | 91.3 | 2.2 | 0.7 | -88.0 | 3.3 | 6.1±1.2 |
| 2X7L-V16D | -40.7 | 18.4 | 2.0 | 39.7 | -1.0 | 19.3±1.0 |
| 3LPH-V16D | -42.7 | 3.8 | 0.3 | 43.4 | 0.7 | 4.8±1.0 |

(a) Positive numbers imply a weakened interaction of the mutated structure interface compared to the wild type.

addition, Lys20 and Tyr23 were mapped as epitope of the multimerisation inhibiting llama single-domain antibody ($Nb_{190}$) and are essential for the higher order multimerisation [27] and in agreement with our computational results. Indeed, it has been shown that $Nb_{190}$ disrupts the higher order Rev multimerisation but not its dimerisation. It is notable that the region comprising Lys20 and Tyr23 shows a strong reduction in RMSF upon binding of individual monomers (Figure 6.2B). As such, Lys20 and Tyr23 have a relatively high entropic penalty, so the inherent flexibility of these amino acids is reduced upon multimerisation of individual monomers at the $\beta$ interface. Of note, conserved residues have been proven to be less flexible in protein-protein interactions [28].

### 6.4.9 Mutational analysis of hot spot residues

To support the binding free energy decomposition explained above, the effect of mutating residues in the interfaces was pursued. A selection of experimentally known mutants were modelled and simulated using the same methodology as on the wild type structures (Table 6.3). All these mutations were introduced by replacing a hydrophobic with a polar residue. The mutations L18T and I55N were applied for investigation of the $\alpha$ interface, while L12S/L60R and V16D were selected to validate the $\beta$ interface. The L12S and L60R mutations were chosen since the 3LPH structure contains these mutations to disrupt the higher order multimerisation site. In addition, L18T is prevalent in HIV-infected patients and the oligomeric assembly is thought to be only partially impaired [21, 22]. No significant entropy contribution differences were found in the previous measurements, therefore these calculations were excluded and shorter production MD simulations (5 ns each) were conducted for efficiency reasons.

The results of these calculations are present in Table 6.3, which displays the relative binding free energy difference of each contribution compared to the wild type sequence. As such, it is possible to compare the strength of each mutation

and to verify which binding free energy contribution changes the most upon mutation. For the $\alpha$ interface, no significant difference was found for the L18T mutation compared with wild type Rev. A minor decrease is found in the relative total electrostatic contribution ($\Delta\Delta G_{ele-tot}$), but this is compensated by slightly stronger apolar interactions ($\Delta\Delta G_{vdw}$ and $\Delta\Delta G_{SA}$). Although threonine is a polar residue, it contains a methyl group, which is able to make beneficial hydrophobic interactions. Intriguingly, the effects of the L18T mutation are very limited. In contrast, a much stronger effect is seen for the I55N mutation due to a loss of both electrostatic and hydrophobic contributions. Stronger effects are also seen for mutations in the $\beta$ interface, where both the L12S/L60R and V16D mutations reduce the affinity between the monomers. These effects are much stronger for the 2X7L crystal structure than for the 3LPH structure. The interface in 2X7L is weaker than the 3LPH interface (see Table 6.1), therefore mutation of a hot spot residue has the strongest effect in the already weakest interaction site.

Evidently, the mutational analysis is flawed to some extent, as the mutations are formed in the complex of Rev monomers, and possibly these complexes are never formed *in vivo*. In fact, many mutations are also known to disrupt the secondary structure of the Rev protein. Nevertheless, these calculations show that a prediction of the relative binding free energies is feasible. Edgcomb *et al.* have ranked the effect of mutating the oligomerisation domains in the following order: V16D > I55N > L60R > L18T > WT [22]. Our measurements correspond with this trend, as the V16D has the strongest effect, L18T the weakest effect, and L60R, L12S, and I55N result in an intermediate disruption. Moreover, it seems that L18T, compared to the other mutations, does not abolish oligomeric binding from a thermodynamical point of view, which is also in line with previous experimental evidence [21, 22].

### 6.4.10 Hydrogen bonds analysis of the interfaces

Hardly any hydrogen bonds involving side chains were present between individual monomers in the simulations due to the hydrophobic nature of the interactions. Hydrogen bond interactions were measured in the last 50 ns of the simulations and studied when more than 20% occupancy was calculated. The few hydrogen bonds found are mainly originating from charged arginine and aspartate residues. In the $\alpha$ interface, hydrogen bonds were found between Arg58 and Ser25 in adjacent monomers and these were present in all simulation systems. Other hydrogen bonds were found only in the 3LPH-AB interaction and thus these interactions are probably not conserved. In the $\beta$ interface, only the Arg17 and Asp9 residues seem to form a stable ionic interaction. This hydrogen bond could be less fixed *in vitro*

because the backbone of Asp9 has been restrained to avoid aspecific interactions between the termini of individual Rev monomers. No hydrogen bonds were found in the $\beta$ interface of the 2X7L system with more than 20% occupancy, although some polar interactions between the Fab and the Rev monomers were present (data not shown).

## 6.5 Discussion

In this chapter we have explored the structural integrity of the Rev multimerisation by performing MD simulations and binding free energy calculations. Recently, two distinct Rev crystal structures have been published. Despite their structural differences, both display a similar behaviour according to our calculations. The MD simulation showed a similar flexibility profile, while the binding free energy calculations indicate quantitatively that the binding interaction and importance of amino acids in each interface is conserved in both crystal structures. The observed hot spot residues are in line with earlier experimental reports [1, 2, 21, 22]. For example, Lys20 and Tyr23 were mapped previously as epitope of the multimerisation inhibiting llama single-domain antibody ($Nb_{190}$) [27] and were identified as key residues using our binding free energy calculations. Furthermore, the calculations suggest additional hot spot residues, such as Lys14, Ser25, Arg48, Gln51, and Arg58 in the $\alpha$ interface and Arg17 in the $\beta$ interface. These residues could be worthwhile for further biophysical investigation. Interestingly, it was shown using a quantitative ranking of the hot spot residues that the residues in the $\alpha$ interface are divided equally over both the N- and C-terminal $\alpha$-helix. In contrast, the most important residues in the $\beta$ interface are positioned on the N-terminal $\alpha$-helix, with only one residue in the C-terminal $\alpha$-helix (Leu60) making a significantly large contribution. The MD simulations also demonstrated that the angle difference between monomers of both crystal structures can adapt to a certain degree. This implies that crystal-packing effects probably cause the reported angle differences between the crystal structures. The variability of the angle also indicates a role of the Rev protein in allowing flexibility at the ARMs to accommodate RNA binding.

## 6.6 Conclusion

Our results have important implications for future drug design efforts from a thermodynamic point of view. Using binding free energy calculations we have shown using a dynamic representation of the Rev protein that the $\alpha$ interface

is much more favourable than the $\beta$ interface. Therefore, the calculated affinity suggests that the dimerisation interface of the Rev protein would be a much more difficult target than the higher order multimerisation. Generally it is challenging to develop small molecule inhibitors, especially against $\alpha$-helical proteins. However, inhibitors against mainly helical interfaces have already been developed, such as the c-Myc/Max inhibitors or PUMA-binding inhibitors [29, 30]. As such, an approach based on the development of small molecule protein-protein interaction inhibitors (SMPPII's) targeted against the Rev multimerisation seems feasible and might allow the generation of a new class of antiviral drugs.

## References


[1] DiMattia, M. A., Watts, N. R., Stahl, S. J., Rader, C., Wingfield, P. T., Stuart, D. I., Steven, A. C., and Grimes, J. M. (2010). Implications of the HIV-1 Rev dimer structure at 3.2 Å resolution for multimeric binding to the Rev response element. *Proc. Natl. Acad. Sci. U. S. A.* 107(13):5810–5814.

[2] Daugherty, M. D., Liu, B., and Frankel, A. D. (2010). Structural basis for cooperative RNA binding and export complex assembly by HIV Rev. *Nat. Struct. Mol. Biol.* 17(11):1337–1342.

[3] Kollman, P. A., Massova, I., Reyes, C., Kuhn, B., Huo, S., Chong, L., Lee, M., Lee, T., Duan, Y., Wang, W., Donini, O., Cieplak, P., Srinivasan, J., Case, D. A., and Cheatham, T. E. (2000). Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc. Chem. Res.* 33(12):889–897.

[4] Onufriev, A., Bashford, D., and Case, D. A. (2004). Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins.* 55(2):383–394.

[5] Andricioaei, I. and Karplus, M. (2001). On the calculation of entropy from covariance matrices of the atomic fluctuations. *J. Chem. Phys.* 115(14):6289–6292.

[6] Carlsson, J. and Aqvist, J. (2005). Absolute and Relative Entropies from Computer Simulation with Applications to Ligand Binding. *J. Phys. Chem. B.* 109(13):6448–6456.

[7] Tintori, C., Veljkovic, N., Veljkovic, V., and Botta, M. (2010). Computational studies of the interaction between the HIV-1 integrase tetramer and the cofactor LEDGF/p75: Insights from molecular dynamics simulations and the Informational spectrum method. *Proteins.* 78(16):3396–3408.

[8] Zhao, Y., Li, W., Zeng, J., Liu, G., and Tang, Y. (2008). Insights into the interactions between HIV-1 integrase and human LEDGF/p75 by molecular dynamics simulation and free energy calculation. *Proteins.* 72(2):635–645.

[9] Christ, F., Voet, A., Marchand, A., Nicolet, S., desimmie, B. A., Marchand, D., Bardiot, D., Van der Veken, N. J., Van remoortel, B., Strelkov, S. V., De Maeyer, M., Chaltin, P., and Debyser, Z. (2010). Rational design of small-molecule inhibitors of the LEDGF/p75-integrase interaction and HIV replication. *Nat. Chem. Biol.* 6(6):442–448.

[10] Cavalluzzo, C., Voet, A., Christ, F., Singh, B. K., Sharma, A., Debyser, Z., Maeyer, M. D., and Eycken, E. V. d. (2012). De novo design of small molecule inhibitors targeting the LEDGF/p75-HIV integrase interaction. *RSC Advances.* 2(3):974–984.

[11] De Luca, L., Barreca, M. L., Ferro, S., Christ, F., Iraci, N., Gitto, R., Monforte, A. M., Debyser, Z., and Chimirri, A. (2009). Pharmacophore-Based Discovery of Small-Molecule Inhibitors of Protein-Protein Interactions between HIV-1 Integrase and Cellular Cofactor LEDGF/p75. *Chem. Med. Chem.* 4(8):1311–1316.

[12] Hess, B., Kutzner, C., van der Spoel, D., and Lindahl, E. (2008). GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput*. 4(3):435–447.

[13] Bussi, G., Donadio, D., and Parrinello, M. (2007). Canonical sampling through velocity rescaling. *J. Chem. Phys*. 126(1):014101.

[14] Parrinello, M. (1981). Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys*. 52(12):7182–7190.

[15] Hornak, V., Okur, A., Rizzo, R., and Simmerling, C. (2006). HIV-1 protease flaps spontaneously open and reclose in molecular dynamics simulations. *Proc. Natl. Acad. Sci. U. S. A*. 103(4):915–920.

[16] Case, D. A., Cheatham, T. E., Darden, T., Gohlke, H., Luo, R., Merz, K. M., Onufriev, A., Simmerling, C., Wang, B., and Woods, R. J. (2005). The Amber biomolecular simulation programs. *J. Comput. Chem*. 26(16):1668–1688.

[17] Hou, T., Wang, J., Li, Y., and Wang, W. (2011). Assessing the Performance of the MM/PBSA and MM/GBSA Methods. 1. The Accuracy of Binding Free Energy Calculations Based on Molecular Dynamics Simulations. *J. Chem. Inf. Model*. 51(1):69–82.

[18] DeLano, W. L. (2002). Unraveling hot spots in binding interfaces: progress and challenges. *Curr. Opin. Struct. Biol*. 12(1):14–20.

[19] Pond, S. J. K., Ridgeway, W. K., Robertson, R., Wang, J., and Millar, D. P. (2009). HIV-1 Rev protein assembles on viral RNA one molecule at a time. *Proc. Natl. Acad. Sci. U. S. A*. 106(5):1404–1408.

[20] Stites, W. E. (1997). Protein-protein interactions: interface structure, binding thermodynamics, and mutational analysis. *Chem. Rev*. 97:1233–1250.

[21] Jain, C. and Belasco, J. G. (2001). Structural model for the cooperative assembly of HIV-1 Rev multimers on the RRE as deduced from analysis of assembly-defective mutants. *Mol. Cell*. 7(3):603–614.

[22] Edgcomb, S. P., Aschrafi, A., Kompfner, E., Williamson, J. R., Gerace, L., and Hennig, M. (2008). Protein structure and oligomerization are important for the formation of export-competent HIV-1 Rev-RRE complexes. *Protein Sci*. 17(3):420–430.

[23] Daelemans, D., Costes, S. V., Cho, E. H., Erwin-Cohen, R. A., Lockett, S., and Pavlakis, G. N. (2004). In vivo HIV-1 Rev multimerization in the nucleolus and cytoplasm identified by fluorescence resonance energy transfer. *J. Biol. Chem*. 279(48):50167–50175.

[24] Wang, J., Morin, P., Wang, W., and Kollman, P. A. (2001). Use of MM-PBSA in Reproducing the Binding Free Energies to HIV-1 RT of TIBO Derivatives and Predicting the Binding Mode to HIV-1 RT of Efavirenz by Docking and MM-PBSA. *J. Am. Chem. Soc*. 123(22):5221–5230.

[25] Huo, S., Massova, I., and Kollman, P. A. (2002). Computational alanine scanning of the 1: 1 human growth hormone–receptor complex. *J. Comput. Chem*. 23(1):15–27.

[26] Zhou, Z. and Madura, J. D. (2004). Relative free energy of binding and binding mode calculations of HIV-1 RT inhibitors based on dock-MM-PB/GS. *Proteins*. 57(3):493–503.

[27] Vercruysse, T., Pardon, E., Vanstreels, E., Steyaert, J., and Daelemans, D. (2010). An Intrabody Based on a Llama Single-domain Antibody Targeting the N-terminal $\alpha$-Helical Multimerization Domain of HIV-1 Rev Prevents Viral Production. *J. Biol. Chem*. 285(28):21768–21780.

[28] Yogurtcu, O. N., Bora Erdemli, S., Nussinov, R., Turkay, M., and Keskin, O. (2008). Restricted Mobility of Conserved Residues in Protein-Protein Interfaces in Molecular Simulations. *Biophys. J*. 94(9):3475–3485.

[29] Mustata, G., Follis, A. V., Hammoudeh, D. I., Metallo, S. J., Wang, H., Prochownik, E. V., Lazo, J. S., and Bahar, I. (2009). Discovery of Novel MycMax Heterodimer Disruptors with a Three-Dimensional Pharmacophore Model. *J. Med. Chem*. 52(5):1247–1250.

[30] Mustata, G., Li, M., Zevola, N., Bakan, A., Zhang, L., Epperly, M., Greenberger, J. S., Yu, J., and Bahar, I. (2011). Development of small-molecule PUMA inhibitors for mitigating radiation-induced cell death. *Curr. Top. Med. Chem*. 11(3):281–290.

# Part IV

# General conclusions

The advent of molecular modelling in biochemical research has spurred novel strategies to investigate protein-protein interactions (PPIs). The increased amount of possibilities can improve our understanding of the interactions between proteins. More importantly, modelling techniques can support rational drug design strategies to seek novel peptidic inhibitors or small molecule protein-protein interactions inhibitors (SMPPIIs).

In this thesis, we have used a combination of MD simulations and binding free energy calculations to decipher the binding interactions between viral proteins and peptides. We believe that such a combination of techniques is essential to achieve success in computer-aided drug design (CADD), as both *conformation* as *affinity prediction problems* need to be tackled. The *conformation problem* states that biomolecules are inherently mobile and therefore a sufficient number of conformations of the molecules of interest should be provided. In an introductory chapter (chapter 2), we have outlined MD simulations as a technique to sample the conformational space of biomolecules. However, a reasonable amount of conformations is not sufficient as such, because an accurate estimation of the predicted binding affinity should be provided as well. This issue, termed *affinity prediction* in biomolecular modelling, depends on the accuracy of the binding free energy method and has been discussed in chapter 3 in detail. The combination of both techniques can be applied to many different fields, such as the two viral targets in this thesis, HIV-1 gp41 FP and Rev (chapter 1).

First, we applied our biomolecular toolbox on the VIRIP:FP interaction (chapter 4). An adapted binding free energy method was proposed and allowed a detailed investigation of the driving forces of the peptide-peptide complex. A number of hot spots were identified that provide hints for further optimisation. Indeed, additional VIRIP derivatives were suggested based on a virtual screening trial. A selection of peptides were tested *in cellulo* and a number were shown to contain improved potency. However, improving the derivatives even further is not straightforward and would possibly require the inclusion of non-classical amino

acids. Alternatively, the peptide backbone can be modified to sterically constrain the conformation of the peptide with its target. The synthesis of these peptides, termed peptidomimetics, is an interesting approach because the configurational entropy of VIRIP was revealed as an important contributor to the binding affinity with gp41 FP. For example, the VIR-165 derivative, which contains an internal disulphide bridge, is sterically more constrained than wild type VIRIP. The reduced entropic penalty in VIR-165 is therefore a partial explanation of the increased potency of this peptide. Next to an increase in potency, the modification of the peptide backbone would also increase the half-life in the human body. However, replacement of the peptide backbone by non-classical scaffolds requires organic chemistry. Considering the many manufacturing steps required for peptidic drug synthesis, such as T20, modifying the peptide backbone would be an expensive and time-consuming task.

Ultimately, it would be more interesting to replace the peptide with an orally administrative small molecule. A rational drug design strategy however depends on an accurate representation of the target of interest. Nevertheless, there have been conflicting reports of the FP conformation in the literature. To this end, we applied multiscale simulations of gp41 FP in solution (chapter 5) and found a highly diverse conformational ensemble in the investigated time scale. To our surprise, we found very few conformations that are proximal to the membrane and VIRIP bound structures. We hypothesize that VIRIP and its derivatives might function by modulating the conformational ensemble, thereby perturbing the fusogenic conformations that co-exist in solution. We can also question how many VIRIP:FP binding modes are present in reality. In contrast to a classical protein-ligand binding case, FP and VIRIP both display a substantial degree of flexibility. Therefore, we propose unrestrained simulations of improved VIRIP derivatives and FP in solution in the future. By simulating both active and inactive peptides, we can perform a full kinetic treatment of the binding process to identify initial binding modes and characterise the most populated binding states. In addition, a number of simulations of the FP in membrane environments will allow a better understanding of the insertion event. We have already performed a number of such studies that will be analysed more in depth in the near future.

We also suggest that the FP might be intrinsically disordered, or at least contains considerably higher mobility compared to standard globular proteins. So-called IDPs are functional proteins that lack a well-defined structure due to a high degree of structural polymorphism. By virtue of the intrinsic disorder of FP, molecular modelling techniques such as MD simulations can be applied to characterise IDP structures, which are usually difficult to obtain accurately from experimental measurements. IDPs are often involved in diseases, such as $\alpha$-synuclein in Parkinson's disease. Those IDPs are however much larger than the viral peptides studied here,

and therefore the *conformation problem* is considerably more complex. To solve the structural ambiguity of IDPs, we propose that the VIRIP:FP complex could be a suitable test case to solve the conformational ensemble of larger IDPs in the future.

The HIV-1 Rev protein is another target of our research (chapter 6). The molecular modelling toolbox was applied to elucidate the binding modes between individual Rev monomers. MD simulations demonstrate substantial fluctuations of the interaxial angles between the monomers. The flexibility of Rev could be essential to recognise multiple RNA binding sites. However, we preferred to study the protein-protein interactions between the Rev monomers and found a number of hot spot residues in each interface. Interestingly, the quantitative ranking of the hot spot binding free energy contributions was in correlation with experimental measurements, which clearly demonstrates the strength of the MM-GBSA method. We also learned that the dimerisation interface is stronger than the higher order multimerisation interface. Therefore, the latter interface would be more interesting to inhibit. Not surprisingly, the nanobody $Nb_{190}$ was identified as a Rev inhibitor by attaching to this higher order multimerisation interface as opposed to the dimerisation interface. Although not discussed in this thesis due to intellectual property protection reasons, a preliminary virtual screening using a 3D pharmacophore model suggested a number of interesting molecules for inhibition of the Rev multimerisation interface. A selection of these molecules from commercial small molecule databases were tested and a number of hit compounds were identified. Possible future experiments could consist of additional virtual screening runs by adjustment of the 3D pharmacophore model based on the active compound alignment. In addition, crystal structures of Rev-$Nb_{190}$ or Rev-compound complexes would give novel clues for the inhibition of Rev multimerisation. Because the Rev protein does not contain a well-defined small molecule binding pocket, the development of effective lead compounds will not be easy. To this end, a number of experimental measurements will be carried out to assess and possibly improve the potency of these antiretroviral compounds in the near future.

Finally, we like to point out that this thesis would not have been possible without the recent advances in computational chemistry. Increases in computer capabilities and the development of improved algorithms have revealed possibilities that were previously unthinkable. Predicting molecular conformations and estimating affinities using binding free energy methods can now be performed much more effectively than just ten years ago. While computational chemistry is not yet able to design drugs "à la carte" against a given target, the amount of possibilities has increased tremendously. Therefore, we believe that this thesis is a clear example of the advantages of rational drug design strategies using computational methods. And it mostly makes us excited of what the future will bring.

# Publications

"These Amazing Simulations
End Up Sounding Even Better
Than The Real Thing"

State Of The Art - Gotye

## Peer reviewed publications

- Venken, T., Krnavek, D., Münch, J., Kirchhoff, F., Henklein, P., De Maeyer, M., and Voet, A. (2011). An optimized MM/PBSA virtual screening approach applied to an HIV-1 gp41 fusion peptide inhibitor. *Proteins*. 79(11):3221–3235
- Venken, T., Daelemans, D., De Maeyer, M., and Voet, A. (2012). Computational investigation of the HIV-1 Rev multimerization using molecular dynamics simulations and binding free energy calculations. *Proteins*. 80(6):1633–1646
- Broos, K., Trekels, M., Jose, R. A., Demeulemeester, J., Vandenbulcke, A., Vandeputte, N., Venken, T., Egle, B., De Borggraeve, W. M., Deckmyn, H., and De Maeyer, M. (2012). Identification of a Small Molecule That Modulates Platelet Glycoprotein Ib-von Willebrand Factor Interaction. *J. Biol. Chem*. 287(12):9461–9472
- Yuan, S., Le Roy, K., Venken, T., Lammens, W., Van den Ende, W., and De Maeyer, M. (2012). pKa Modulation of the Acid/Base Catalyst within GH32 and GH68: A Role in Substrate/Inhibitor Specificity? *PLoS One*. 7(5):e37453
- Deckers*, S. M., Venken*, T., Khalesi*, M., Gebruers, K., Baggerman, G., Lorgouilloux, Y., Shokribousjein, Z., Ilberg, V., Schonberger, C., and Titze, J. (2012). Combined Modeling and Biophysical Characterisation of CO2 Interaction with Class II Hydrophobins: New Insight into the Mechanism Underpinning Primary Gushing. *J. Am. Soc. Brew. Chem*. 70(4):249–256 (* shared first author)

- Khalesi, M., Venken, T., Deckers, S., Winterburn, J., Shokribousjein, Z., Gebruers, K., Verachtert, H., Delcour, J., Martin, P., and Derdelinckx, G. (2013). A novel method for hydrophobin extraction using CO2 foam fractionation system. *Industrial Crops & Products*. 43:372–377

- Vercruysse, T., Boons, E., Venken, T., Vanstreels, E., Voet, A., Steyaert, J., De Maeyer, M., and Daelemans, D. (2013). Mapping the Binding Interface between an HIV-1 Inhibiting Intrabody and the Viral Protein Rev. *PLoS One*. 8(4):e60259

- Venken, T., Voet, A., De Maeyer, M., De Fabritiis, G., and Sadiq, S. K. (2013). Rapid Conformational Fluctuations of Disordered HIV-1 Fusion Peptide in Solution. *J. Chem. Theory Comput*. 9(7):2870–2874

## Submitted publications

- Venken*, T., Van Eyck*, D., De Raeymaecker, J., Voet, A., and De Maeyer, M. GROMACS File Processor, a tool to facilitate Molecular Dynamics analysis (* shared first author)

## Publications in preparation

- Venken, T., Voet, A., and De Maeyer, M. Molecular dynamics simulations applied in HIV-1 drug discovery, the envelope as target

- Venken, T., Voet, A., Zhang, K. Y. J., and De Maeyer, M. Molecular dynamics simulations of HIV-1 gp41 fusion peptide in membranes: the influence of cholesterol

## Lectures given at international conferences

- **Binding free energy calculations on a gp41-FP HIV-1 inhibitor**
  NBIC conference: Mutations in proteins: structure, function, dynamics and disease
  Nijmegen, The Netherlands, 20/09/2010
- **Investigating viral proteins with MD simulations**
  Technical Meeting on High-Troughput MD
  Barcelona, Spain, 08/11/2013

## Poster presentations given at international conferences

- **An optimized MM/PBSA virtual screening approach applied on an HIV-1 gp41 fusion peptide inhibitor**
  8th EBSA European Biophysics Congress
  Budapest, Hungary, 23–27/08/2011
- **Virtual screening of an HIV-1 gp41 fusion peptide inhibitor**
  Annual One-Day Meeting on Medicinal Chemistry of SRC & KVCV (MedChem 2011)
  Ghent, Belgium, 25/11/2011

## Awards

- **Structural analysis of the protein databank for the presence of alternative tryptophan conformations.**
  GOA midterm meeting of Multicentre quantum chemistry
  Blanden, Belgium, 18/11/2010.
  *Best young speaker award*
- **The HIV-1 gp41 fusion peptide as novel antiviral target.**
  11th Chemistry Conference for Young Scientists (ChemCYS),
  Blankenberge, Belgium, 1-2/03/2012.
  *Best oral presentation in the field of Biochemistry & Medicinal Chemistry*

FACULTY OF SCIENCE
DEPARTMENT OF CHEMISTRY
LABORATORY FOR BIOMOLECULAR MODELLING
Celestijnenlaan 200G, box 2403
B-3001 Heverlee
tom.venken@chem.kuleuven.be
http://chem.kuleuven.be/en/research/bmsb/biomol